Department IV
Data Science
Winter Term 2023

Interactions of Variable Selection and Inference

Supervisor:
Caratiola,Christopher

Submitted on:
31 March 2024

Submitted by:
Shubham Tiwari
Aditya Neekhra
Pushpamala Aralakuppe Rajeev
Sargur Abhinandan Kengegowda

Contents

List of Figures

List of Tables

1 Introduction

In statistical modeling, variable selection and inference are essential procedures that improve the results' interpretability, precision, and generalizability. Finding the most pertinent predictors from a wider pool of possible variables to include in a statistical model is known as variable selection or variable selection in machine learning. Eliminating unnecessary or redundant variables will help the model perform better by lowering computational complexity, reducing overfitting, and improving interpretability(1).

However, inference is the process of drawing conclusions about a population from sample data. Inference is helpful in determining the relationship between the selected variables and the intended outcome when it comes to variable selection. **Our research focuses on**

1. **Evaluating Variable Selection Methods for Unbiased Predictors:** How do traditional variable selection methods fare in accurately identifying significant uncorrelated predictors amidst a backdrop of correlated predictors and inherent noise within datasets?
2. **Challenges in Crafting and Utilizing Synthetic Data:** What obstacles and constraints arise in the creation of synthetic datasets designed to mirror the intricacies and subtleties of authentic data scenarios, and what is their impact on the formulation and validation of predictive models? The utility of synthetic datasets is undeniable in the stages of model development, testing, and validation. Nonetheless, capturing the stochastic behaviors and interactions of real-world data within synthetic datasets is inherently challenging. This line of inquiry seeks to bridge the disconnect between the controlled environment of synthetic data and the complex reality it attempts to replicate.

Using the right variable selection techniques throughout the model construction phase is essential, regardless of the modeling approach employed.Which variables to include in a model is often seen as the most important and difficult phase in the process of creating models. Numerous advantages come with variable selection, including enhanced prediction performance of the models, faster and more economical delivery of variables by cutting down on training and utilisation time, easier data visualisation, and a generalised improved comprehension of the underlying process that produced the data(2).

There are lot of advantages of varaible selection, such as reducing the requirement for measurement and storage, accelerating training and application, escaping the dimensionality curse to improve prediction accuracy, and enhancing data visualization and comprehension. One further way that this special issue differs from earlier research is that certain approaches emphasise some aspects more than others.

The selection and generation of variable subsets that are helpful for creating effective predictors is the main emphasis of the majority of the works in this area. On the other hand, it might be difficult to determine which variables are important and to rank them.In most cases, especially when there are duplicate variables, it is not optimal to choose the most important variables when developing a predictor.Nevertheless, a list of helpful factors could exclude some important but superfluous variables. (2).

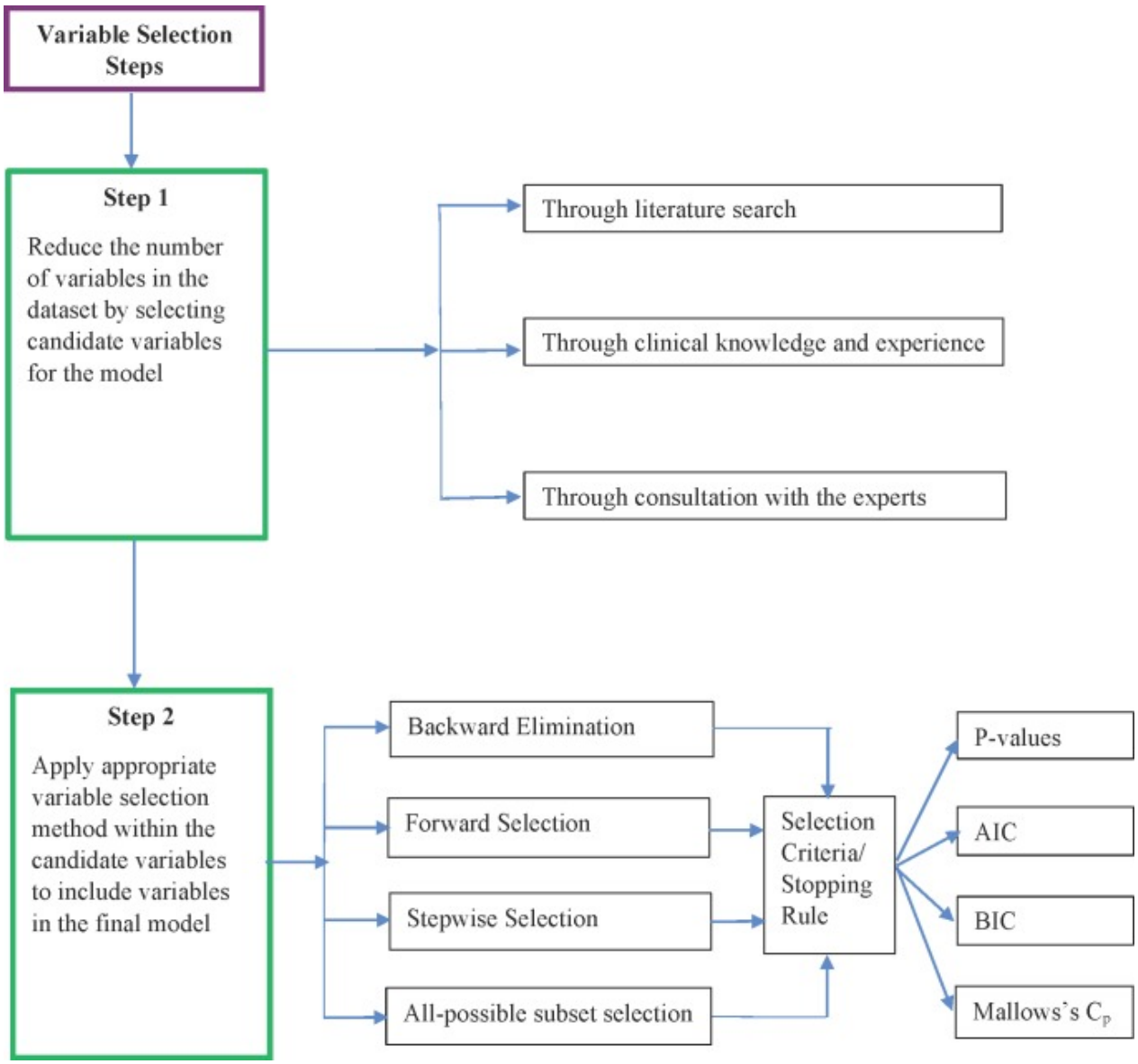

Figure 1: Variable Selection

2 Literature Review

The variable selection and inference processes of statistical models are essential for comprehending and forecasting complicated occurrences in a variety of scientific domains. To create precise, understandable, and broadly applicable models, it is imperative to comprehend the complex interrelationships between these systems. A succinct summary of the major advancements, difficulties, and patterns in this field is given in this literature review(4).

Variable selection is critical to model optimization and improved interpretability, as previous studies have shown. Using statistical criteria, stepwise regression is a classic approach in this search that adds or removes variables one after another. But because of their propensity to increase bias and lower mistake rates, these techniques were frequently criticized(5).

Significant progress has been achieved since the introduction of regularisation methods such as LASSO and Ridge Regression. By adding penalty clauses to loss functions, these methods effectively perform a variable selection during the estimation of coefficients in order to reduce their size. By combining Ridge and LASSO sanctions, the Elastic Net was able to improve upon this method for addressing multicollinearity and variability of selection problems in high dimensional datasets.(6)(7).

Moreover, since it uses probability distribution to describe known knowledge and uncertainty of model parameters, the Bayesian framework provides improved methods for selecting and evaluating variables. At the most basic level, this paradigm allows for a better understanding of model uncertainty and variability in importance.(8).

In order to enhance variable selection and inference, recent work has concentrated on fusing conventional statistical techniques with machine learning methodologies. The range of tools accessible to researchers is expanded by methods such as Random Forests and Gradient Boosting Machines, which yield varying significance measures from intricate and non-linear models(9).

In spite of these developments, problems nevertheless persist. Other ongoing research concerns include the trade-off between interpretability and model complexity, managing correlated predictors, and developing resilient techniques to different data formats and architectures. Scalable and computationally efficient solutions are becoming increasingly important as large datasets become more widely available

Finally, the requirement for reliable, comprehensible, and predictive models drives ongoing study into variable selection and inference. Further study on variable selection in relation to causal inference and the development of more adaptable algorithms including different kinds of data are possible avenues for the future(10).

3 Methods

An important step in the modeling process is variable selection, which allows you to choose the most important predictors from a potentially enormous number of variables. This procedure helps to decrease computational complexity and improve interpretability in addition to increasing the model's performance. Numerous techniques have been devised for the purpose of variable selection, each with advantages and disadvantages. Below is a summary of some of the most popular methods.

3.1 Filter Methods

A class of variable selection strategies called filter methods depends on the inherent characteristics of the data. Using statistical measurements, they evaluate the significance of variables and choose variables apart from the model. Frequently utilized standards in filtering techniques consist of.

3.1.1 Correlation Coefficient

Two or more variables have a linear connection, which may be measured using correlation. The ability to anticipate one characteristic based on another is known as correlation. A strong correlation between the goal and desirable variables is the rationale for utilizing correlation for variable selection. In addition, variables ought to be uncorrelated among themselves but connected with the aim instead. One variable can be predicted from the other if there is a correlation between the two. Due to the fact that the second variable does not provide more information, if two characteristics are linked, the model just requires the first. Pearson Correlation is primarily used here(11).

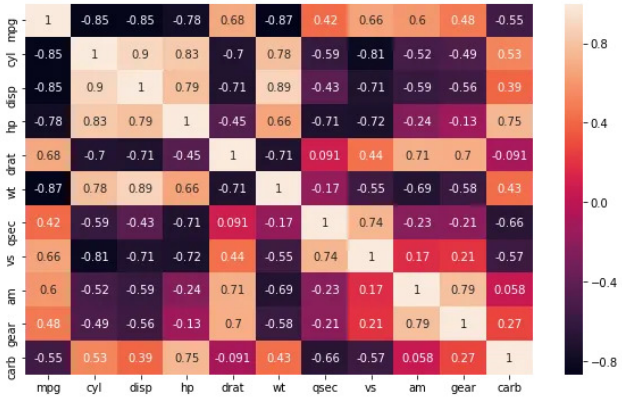

Figure 2: Correlation Coefficient

3.1.2 Chi-square test

The Chi-square test is used to a dataset's categorical variables. Calculating the Chi-square between each variable and the target allows us to find the minimum number of variables that must have the greatest Chi-square scores. The chisquared test may be used to assess the association between various dataset variables and the target variable only if the variables are categorical, sampled independently, and have values with an expected frequency greater than 5. (11).

3.1.3 Information gain

Information Gain, also called Mutual Information, assists in measuring the degree of dependence between the two variables. Mutual information intuitively quantifies the knowledge that X and Y share, i.e., the extent to which understanding one of these variables lowers uncertainty regarding the other. Their mutual information is 0, for instance, if X and Y are independent. This means that knowing X does not provide any knowledge about Y, and vice versa. However,When X and Y are deterministic functions of one another, then X shares all of its information with Y; understanding X influences the value of Y and vice versa.

3.2 Wrapper Methods

With wrapper methods, choosing a collection of variables is viewed as a search issue in which several combinations are created, assessed, and contrasted with one another. These methods use a predictive model to score variable subsets and select the one that yields the best performance according to a predefined criterion(11).

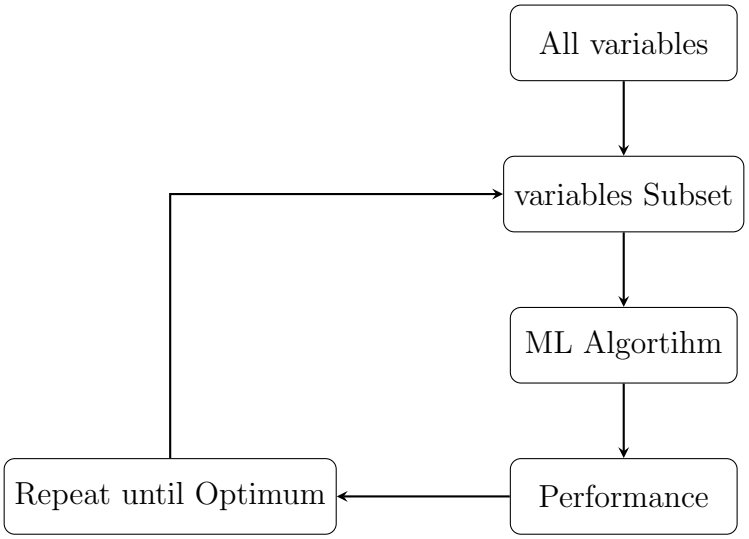

Figure 3: Flowchart for Wrapper Methods

3.2.1 Forward selection

In statistical modeling , forward selection is a popular and straightforward technique that is especially helpful for datasets with a lot of predictors used or variable selection. This method is categorized as a wrapper method, which is more general and evaluates subsets of variables by fitting models and determining their predictive potential. Selecting the variable that best fits the model at each step based on a predetermined criterion, forward selection begins with the simplest model and proceeds one variable at a time until no appreciable improvement is possible. The converse of the backward elimination method is the forward selection strategy. We add variables one at a time until we are unable to identify any that improve the model, as opposed to removing them one at a time. The forward selection process is as follows.

- **Initialization:** Start with a model that has no predictors at all. When all predictor variables are removed from the model, the intercept which stands for the mean of the response variable is the sole component remaining in the model.
- **Model Evaluation Criterion:** In the context of hypothesis testing, often utilized criteria consist of the adjusted R-squared, the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC), and the p-value. The model's objective determines which criterion to use (e.g., prediction accuracy vs. interpretability).
- **Variable Addition:**
  - Think about including every variable that isn't already in the model at each step.
  - A new model is fitted for each candidate variable, adding the candidate to the previously chosen variables. The selected criterion is used to evaluate the improvement in the model fit.
  - The variable (e.g., the lowest AIC or BIC, the highest adjusted R-squared, or the most statistically significant p-value) is included to the model whose inclusion results in the most significant improvement.
  - One variable is included at a time in this manner until the model meets the selected criterion without any more variables showing a discernible improvement.
- **Stopping Criteria:**When the model's fit is not considerably improved by adding any of the remaining variables, the process comes to an end. The standards for "significant improvement" can take many forms, but they typically entail either a small improvement in AIC, BIC, or adjusted R-squared, or establishing a threshold for the p-value (a variable is added only if its inclusion results in a p-value below 0.05)(12).

3.2.2 Backward elimination

The whole model, with all independent effects included, is where the backward elimination approach begins. After then, impacts are eliminated one at a time until a stopping condition is met. The impact that contributes the least to the model is eliminated at each phase. An F statistic is used in conventional backward elimination implementations to evaluate an effect's contribution to the model. The predictor with the least significant F statistic is discarded after the process is repeated until all effects in the model have F statistics significant at a stay significance level (SLS)(13).

$$F = \frac{(RSS_{p-k} - RSS_p)/k}{RSS_p/(n-p-1)} \tag{1}$$

Steps in Backward Elimination is as follows.

1. Choosing a significance level or P-value is the first and easiest step in the backward elimination process. A significance threshold of 5% is often chosen in most situations. Thus, 0.05 will be the P-value. Depending on the project, you can alter this value.

2. Just put all the specified characteristics into your machine-learning model. Consequently, you incorporate each of the 100 characteristics into your model and fit it to your test dataset. This is unchanged.

3. variable or predictor with highest P-value are identified.

4. It is one of the important step. We make decisions at this point. Earlier, the variable with the highest P-value was identified. When a characteristic's P-value exceeds the significance level we selected in the first stage, it is eliminated from our dataset.If the P-value of the largest variable in the set, this variable, is less than the significance threshold, then we will move on to Step 6, signaling that we are done. If the highest P-value surpasses the significance threshold, remember to remove the variable.

5. Once the variable has been found, this phase will include eliminating it from the dataset. We remove the variable from the dataset so that we may use the new dataset to fit the model. After fitting the model to the new dataset, we'll go back to step 3.Until step 4, when the significance selected in step 1 is less than the largest P-value from all the variables remaining available in the dataset, this process is repeated.

6. The variable selection procedure is completed once we reach step 6. We were able to eliminate elements that weren't important enough for our model by employing backward elimination(13).

3.2.3 Stepwise regression

Stepwise regression is the iterative, step-by-step process of creating a regression model while choosing which independent variables to include in the final model.It involves progressively including or removing likely explanatory variables, with a statistical significance test at the conclusion of each cycle. Stepwise regression is a statistical method for fitting regression models where the predictive variables are chosen automatically. In each step, a variable is added to or deleted from the set of explanatory variables based on a predefined standard.This often takes the shape of an F-test or t-test sequence that is run forward, backward, or combination(14). The following crucial phases are included in the stepwise regression technique, which alternates between adding and deleting variables based on predetermined criteria.

1. **Initialization:** As in forward selection, begin with no variables in the model, or as in backward elimination, begin with all candidate variables. Establish a threshold of relevance for both joining and remaining in the model. 0.05 for entering (α_{enter}) and 0.10 for removal (α_{stay}) are typical selections.

2. **Model Evaluation:** Using p-values from t-tests, determine the statistical significance of each variable inside the model (for backward exclusion) or outside the model (for forward selection).

3. **Variable Addition (Forward Selection Step):** Choose the variable that most enhances the model from those not yet included in it based on a predetermined criterion (often the lowest p-value less than α_{enter}). To the model, add this variable.

4. **The Variable Removal (Backward Elimination Step):** involves reevaluating every variable present in the model once a new variable is introduced. A variable is eliminated from the model if it does not satisfy the requirement for remaining in the model, which is having a p-value bigger than α_{stay}.To make sure that the addition of additional variables doesn't make the previously chosen variables less important, this step is essential.

5. **Iteration:** Until no more variables can be added or deleted in accordance with the entry and stay conditions, repeat the addition and removal stages. Until a model is achieved where all variables are significant and no significant variables are outside the model, this iterative procedure is continued.

6. **Model Finalisation:** At the conclusion of the iterative process, a final model is produced that includes variables that were shown to be highly significant predictors of the outcome variable(15).

3.3 Embedded Methods

During training, embedded techniques pick variables that are unique to individual models.These methods integrate the selection process into the model fitting process and can be more efficient than filter or wrapper methods. The selection of variables is integrated into the model's training process using embedded methods of variable selection. Variable selection is done directly into the model fitting process in embedded methods, as opposed to filter or wrapper methods, which handle it as a separate step either before or after model training. Because of this integration, embedded approaches are frequently more effective and can provide models that are easier to understand and comprehend, particularly when working with high-dimensional data. Important variables of Embedded Techniques.

- **Integration:** Variable selection is not an independent phase, but rather a crucial component of the model training procedure.
- **Efficiency:** Embedded techniques have the potential to be computationally more efficient than wrapper methods by merging the phases of variable selection and model training.
- **Regularisation:** To penalize the model's complexity, several embedded approaches include regularisation techniques, which inevitably lead to variable selection.

Some Embedded Methods commonly used are

3.3.1 LASSO

Gives the regression model a penalty term that limits the total of the model coefficients absolute values. Variable selection is accomplished by the penalty's ability to reduce certain coefficients to zero.The statistical model's interpretability and prediction performance are enhanced by variable selection and regularization using the LASSO regression analysis technique. Assuming that there are comparatively few non-zero coefficients in the linear model, the LASSO method functions on the sparse assumption.

Lasso regularisation is the process or method of incorporating a penalty into the regression model in order to promote sparsity (simplicity) and prevent overfitting. In contrast, Lasso shrinkage characterizes the result of this regularisation, which is the decrease in coefficient sizes, sometimes to zero, resulting in more straightforward and understandable models(16)

Due to its ability to decrease the size of the independent variable coefficients based on their predictive strength, LASSO is an effective shrinkage estimator. We may limit the model to variables having coefficients that are not zero by allowing some of the coefficients to decrease to zero(17).

A family of penalized least squares estimators includes methods such as ridge regression and elastic net, of which LASSO is only one member. Initially focusing on the cost function of linear regression:

$$J = \frac{1}{2m} \sum_{i=1}^{m} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j \right) \tag{2}$$

In this linear regression framework, the dependent variable is denoted by y_i, while x_{ij} represents the independent variables. The coefficients, represented by β_j, are determined by minimizing the cost function J. With m being the total number of observations and p the count of predictors, the goal is to find the optimal coefficients that lead to the best fit. Overfitting a scenario where the model fails to generalize to new data is a common problem in such models. A LASSO model, which incorporates a penalty term, can be used to mitigate this issue by introducing regularization.
One way to address this issue if the linear regression model is overfit and underpredicts incoming data is to build a LASSO model with a penalty term:

$$J = \frac{1}{2m} \sum_{i=1}^{m} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right) + \lambda \sum_{j=1}^{p} |\beta_j| \tag{3}$$

where the parameters are the same as before with the addition of the regularization parameter λ. By adding these extra terms the cost of $beta_j$ is increased have to be reduced.A lower value of $beta_j$ will "smooth out" the function, making it fit the data less closely and increasing the likelihood that it will generalize well to fresh data. The amount that the cost of $beta_j$ is raised is determined by the regularisation parameter λ. With cross-validation, a data-driven technique for determining λ based on its predictive power, LASSO estimation in EViews may automatically choose an acceptable value(18).

Variable selection using the LASSO shrinkage regression model.

Let $y_i = 0$ or 1 denote the binary outcome of the ith sample of n individuals. The probability of observing $y_i = 1$ is written as $p_i = Pr(y_i = 1)$, $i = 1, ..., n$. Let $\mathbf{x}_i = (x_{i1}, ..., x_{ip})^T$ denote a p-dimensional vector of predictors that may be associated with the outcome y_i. We model the relationship of y_i with $\mathbf{x}_i$ through a logistic regression model:

$$logit(p_i) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}, \tag{4}$$

where β_0 and β are the intercept and regression coefficients, and $\text{logit}(p_i)$ is estimated by:

$$logit(p_i) = \log \left(\frac{Pr(y_i = 1)}{1 - Pr(y_i = 1)} \right), \tag{5}$$

We define the parameter vector $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)^T$, and then the above logistic regression model may be expressed in a condensed way as: $\text{logit}(p_i) = \mathbf{x}_i^T \boldsymbol{\beta}$.
The LASSO estimates $\hat{\boldsymbol{\beta}}$ are given by:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left(\sum_{i=1}^{n} \left(y_i - \mathbf{x}_i^T \boldsymbol{\beta} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right), \tag{6}$$

where the tuning parameter, λ, is non-negative. The penalty function in LASSO con-

stantly decreases the coefficients toward zero; the more λ is shrunk, the more shrinkage occurs. A sparse subset of variables with non-zero regression coefficients will be obtained by the LASSO model as λ grows, as some of the coefficients will decrease to exactly zero(20).

Important Characteristics of LASSO

- **Variable Selection:** LASSO selects variables by making some of the coefficients precisely zero. By preventing overfitting, this can result in models that are simpler to understand and frequently perform better on fresh, unobserved data.
- **Sparsity:** LASSO often yields sparse models, particularly when there is a high level of multicollinearity among the predictor variables or when the number of predictors is significantly more than the number of data.
- **Regularization:** The coefficient of regularization, denoted by λ, plays a pivotal role in model complexity. In the context of LASSO regression, setting λ equal to zero effectively transforms the method into an ordinary least squares regression that includes every variable. As the value of λ is increased, the model enforces greater sparsity by driving more coefficients towards zero, thereby streamlining the model. The process of cross-validation is often employed to ascertain the most appropriate value for λ(21).

3.3.2 Decision Trees

Decision trees are a non-parametric supervised learning method for regression and classification. They may also be used successfully for variable selection because of their innate capacity to divide data into subsets according to the significance of individual variables. When creating a decision tree, one must decide which variables or characteristics to divide on at each level of the tree. This procedure may give a basic overview of the relative importance of each variable.

Steps of Decision tree for Variable Selection

- **Splitting Binary Recursively:**The decision tree approach starts at the root of the tree and iteratively explores every variable (and all potential values of that variable) to determine which split would minimise the loss function, which is often quantified by mean squared error for regression and Gini impurity for classification.
- **variable Importance Scoring:** The decrease in the loss function for each variable is calculated each time it is utilised to split the data. A characteristic is seen more significant the more it lessens the loss. This statistic is frequently normalised to ensure that the total significance ratings for all characteristics add up to one.
- **Pruning trees:**Pruning trees helps prevent overfitting, although it may have an impact on variable significance ratings. A more unified model may be produced by pruning, which can eliminate branches that are not very important.

- **Variable Selection:** The characteristics that are employed in the tree can be deemed significant after it is constructed (and perhaps trimmed), with a ranking determined by the relevance scores assigned to each variable(22)(23).

3.4 Correction Methods

3.4.1 Residual Bootstrap

The residual bootstrap is a resampling approach used in statistics that is mostly used to sample with replacement from a model's residuals estimating the distribution of a test statistic or estimator. When the estimator's theoretical distribution is unclear or complicated, the approach is very helpful. The residual bootstrap may be used to evaluate and enhance the reliability of variable selection processes or model coefficients, even though it is not a variable selection or correction method in and of itself(27).

Procedure for Residual Bootstrap

1. **Model Estimation:** Begin by estimating a regression model using the given dataset to calculate both the coefficients and the residuals, which represent the discrepancies between the actual and model-predicted values.

2. **Residual Sampling:** Perform a random sampling of the residuals with replacement to construct a bootstrap sample, ensuring this sample is of equivalent size to the initial residuals array.

3. **Bootstrap Response Construction:** Create a synthetic response variable through the addition of the bootstrapped residuals to the predicted values from the original regression equation.

4. **Model Re-estimation:** Employ the initial set of predictors alongside the synthetic response variable to recalibrate the regression model and procure a fresh batch of coefficient estimations.

5. **Iterative Repetition:** Execute the above steps (2 to 4) repeatedly, typically around 1,000 to 10,000 iterations, to generate a robust distribution for the coefficient estimates(28)(29).

3.4.2 Restricted Permutation

In order to account for the structure or constraints present in the data, this statistical approach is mostly utilised in the context of hypothesis testing, especially in permutation tests. To make sense of it, let's first explain the idea and then apply it to data analysis, which is arguably connected to taking the variable structure into account or adjusting for it.
Instead of randomly rearranging the whole dataset, restricted permutation includes doing so

within selected blocks or groups that share specific attributes. The groups' integrity or the data's structure, which may be crucial for the analysis, is preserved using this technique.

Important Procedures in Restricted Permutation

1. **Establishing Constraints:** Constraints are established based on the specific design of the study or the intrinsic structure of the dataset. For instance, within a clinical study, stratification of patients may occur according to variables such as age, gender, or severity of condition. Similarly, in studies of an environmental nature, grouping may be based on locations or intervals of time.

2. **Execution of Permutations:** Within the confines of each distinct group established by the aforementioned constraints, elements of the data (such as the allocation of treatments or the outcomes measured) are subjected to random permutation repeatedly. This procedure ensures the maintenance of the group's structural integrity and concurrently generates a null distribution for hypothesis testing that posits no relational or effective differences.

3. **Derivation of Test Statistics:** A specific test statistic is derived from each permutation process, such as the mean differential between groups undergoing different treatments. The collation of these test statistics, obtained from extensive permutation, provides a referential foundation to ascertain the statistical significance of the test statistic observed from actual data.

Restricted permutation is important in the analytical process since it helps comprehend the influence of variables while taking data structure and constraints into consideration. However, it does not itself fix variables. Restricted permutation provides an effective means of rigorously testing hypotheses without going against the basic assumptions associated with a data structure in situations where the data structure may add bias or confounding into the research(30).

3.4.3 Resampling with Replacement

A statistical technique called bootstrapping, or resampling with replacement, involves taking multiple samples from the data in order to enhance the estimation of the sampling distribution of a statistic. Although bootstrapping is not a variable selection method in and of itself, it may be used in conjunction with techniques such as bootstrap aggregating (bagging) to improve variable selection methods or evaluate model stability or variable relevance.

1. **Creation of Bootstrap Samples:** Commence by generating multiple bootstrap datasets from the original data. Each of these datasets should match the original in size and be constructed by randomly choosing observations with the possibility of repetition.

2. **Model Application:** Apply the prediction model to every individual bootstrap dataset. This technique is prevalent in a range of models that are often paired with bootstrapping for variable selection, such as linear and logistic regression, as well as decision tree models.

3. **Assessment of Variable Significance:** Post model fitting, appraise the significance of the variables in the context of the bootstrap dataset. Evaluation metrics vary and are dependent on the model type for instance, the magnitude of the regression coefficients or the Gini importance in decision trees.

4. **Synthesis of Findings:** Conclude the process by executing the resample and model application steps repetitively, followed by an amalgamation of the significance scores of the variables from all bootstrap iterations to gauge their overall impact(31).

Difference between Variable Selection,Potential Estimation and Correction methods

Variable Selection is about choosing which variables to include in a model. It's a critical step in model design that can significantly impact performance and interpretability. Parameter Estimation deals with finding the most accurate values for the coefficients of the variables included in the model. This process is about quantifying the strength and direction of relationships between predictors and the outcome. Correction Methods are employed to refine models or estimates, making adjustments to overcome specific problems or limitations in the modeling process. They may be used in a number of ways to guarantee the validity and robustness of the results, such as before, during, or after parameter estimation and variable selection(22).

4 Dataset Challenge

Purpose of Synthetic Dataset Generation

The synthetic dataset is crafted to challenge and test the robustness of variable selection methods and the predictive capabilities of models. The dataset simulates common data complexities including:

- **Heterogeneous variable Correlations:** variables possess different levels of linear or non-linear relationships among themselves and with the target variable.
- **Signal and Noise Differentiation:** The dataset includes both signal-carrying and noise-inducing variables, demanding sophisticated variable selection to identify truly predictive attributes.
- **Realistic Data Irregularities:** By adding noise, we replicate real-world imperfections such as measurement errors or unexplained variability.

Description of variables

- **Base variable (x1):** The cornerstone of the synthetic dataset, x_1 is generated to have a uniform distribution across a specific range, which is the starting point for creating correlated variables.
- **Correlated variables (x2, x3):** These are engineered by adding Gaussian noise to x_1. x_2 maintains the direct relationship with x_1, while x_3 is designed to have an inverse relationship, adding a layer of complexity.
- **Derived Correlated variables (x4, x5):** These variables are constructed to have indirect relationships with x_1 through x_2 and x_3, respectively. They incorporate additional random noise, which simulates the imperfect correlations typically seen in real datasets.
- **Uncorrelated Noise variables (z1 to z4):** These are independent of the correlated variables and are included to represent variables that do not hold intrinsic predictive power for the target variable y.
- **Target Variable (y):** With noise added, it is created by combining linearly correlated and uncorrelated data. This combination simulates the composite character of real-world data responses by weighting each component to represent its varied effect on the result.

Importance for Research

In machine learning research, this synthetic dataset fulfills a number of purposes:

- **Model Evaluation:** Provides a controlled environment for determining the accuracy and resilience of predictive models.
- **variable Selection Analysis:**Investigate how different variable selection strategies discover and prioritize correlated versus uncorrelated data.
- **Algorithmic Comparison:**Enables R comparative analysis of several algorithms, particularly their capacity to manage multicollinearity and variable redundancy.
- **Education Application:**Works as a priceless teaching tool, helping newcomers to the profession and students alike comprehend the complexities of dataset properties and how they affect model training and assessment.

5 Implementation

5.1 LASSO

Model

In our research we have used Lasso Regression technique to find important predictors and we have ignored those which are not important for our model.Our method is a strict preprocessing pipeline which ensures consistency in scaling by standardising variables.

The scikit-learn StandardScaler centers values on zero with unit variance. We applied this preprocessing step. It's crucial when regularization methods, like LASSO, penalize coefficients. Otherwise, larger scales would distort variable importances. Standardization equalizes variables before modeling. It avoids bias from differing units and magnitudes.

We utilized LASSOCV after scaling. It's a LASSO regression version with cross-validation to find the optimal regularization parameter, alpha. Avoiding overfitting through this process is key for the model to generalize to new, untested data. We employed a 5-fold cross-validation approach, striking a balance between robust model validation and computational efficiency.

LASSO regression allows us to identify key variables. It assigns non-zero coefficients to important ones. variables with zero coefficients are deemed unimportant. Removing them won't impact predictions much. This model excels at variable selection. It's useful for complex datasets with many dimensions. Such data can pose difficulties known as the "curse of dimensionality." Our findings showcase LASSO's strength in this area.

LASSO regression offers understanding of data structure. It separates informative traits from unnecessary or redundant ones. LASSO regression is a variable selection technique. But it's also a tool that helps understand how data was created

The model we get after selecting variables shows how much it relies on fewer predictors. This simpler model makes it easier to understand how each important variable affects the target variable. This highlights how well the LASSO regression improves interpretability and transparency something vital for people who depend on the model's results.

We can gain a better grasp of the connections in our data through close study of the key and unimportant aspects uncovered here. The less crucial ones help reevaluate data collection techniques and cleaning methods that could boost model performance. Meanwhile, the vital traits deserve deeper analysis regarding their ties to the target variable.

Algorithm 1 Restricted Permutation Test Algorithm

```
threshold ← 0.01
lasso_significant_vars ← ColumnsGreaterThan(X, lasso.coef, threshold)
lasso_nonsignificant_vars ← ColumnsLessOrEqualTo(X, lasso.coef, threshold)
type_1_errors ← InitDictionary(lasso_nonsignificant_vars)
type_2_errors ← InitDictionary(lasso_significant_vars)
n_permutations ← 10
for i ∈ {1 ... n_permutations} do
    X_permuted ← PermuteNonSignificant(X, lasso_nonsignificant_vars)
    X_combined ← Concatenate(X[lasso_significant_vars], X_permuted)
    X_permuted_scaled ← Standardize(X_permuted)
    FitModel(lasso, X_combined, y)
    for all var ∈ lasso_nonsignificant_vars do
        if CoefNonZero(lasso.coef, var) then
            type_1_errors[var] ← type_1_errors[var] + 1
        end if
    end for
    for all var ∈ lasso_significant_vars do
        if CoefZero(lasso.coef, var) then
            type_2_errors[var] ← type_2_errors[var] + 1
        end if
    end for
end for
```

Restricted Permutation

Bootstrapping

- **Consistency of Results:** A predefined random seed guarantees that findings can be replicated, bolstering the research's scientific integrity.
- **Resampling with Bootstrap:** Multiple bootstrap datasets derived from the residuals of an initial LASSO model are generated to scrutinize the consistency and dependability of selecting variables.
- **Monitoring Errors:** The method numerically evaluates Type 1 and Type 2 errors throughout the bootstrapping iterations, offering insights into the frequency of incorrectly flagged significant variables (Type 1) and overlooked significant variables (Type 2).
- **Evaluating LASSO's Reliability:** The bootstrap approach is utilized to gauge the steadiness of the LASSO model's variable selection against variations in the sample.
- **Assessment of Error Probabilities:** Calculating error rates is crucial for gauging the potential for incorrect variable identification or missed variables within the context of variable selection.
- **Influence on Model Interpretation:** The conducted analysis sheds light on the robustness of the LASSO regression, aiding researchers in accurate model interpretation and in establishing the validity of their variable selection(33).

5.2 Forward Selection

Model

- **Stepwise variable Inclusion:** This algorithm add one variable into a model at a time.It gives preferences to those p-values who have minimal p-values in regression outcomes.
- **Criterion Based on Statistical Significance:** Establishing a p-value threshold at 0.05 is the accepted practice. This criterion governs which variables qualify for inclusion in the model. It aims to curb the likelihood of falsely rejecting a true null hypothesis, an error known as Type I.
- **Evaluation of P-values:** Within the iterative process, each potential variable is evaluated for its contribution by integrating it into the existing model and performing regression analysis, selecting the variable with the most statistically significant p-value for addition.

- **Prevention of Overfitting:** To avoid the inclusion of superfluous variables, the variable selection ceases once all remaining variables exhibit p-values above the established significance threshold, thus safeguarding the model against overfitting.
- **Construction of the Model:** The final product is a regression model endowed with the optimally chosen variables, ready for subsequent examination and inference.
- **Benefits for Model Parsimony and Effectiveness:** The deliberate inclusion of only significant variables yields a model that is both easier to interpret and potentially more robust, steering clear of the pitfalls associated with overly complex models.

Restricted Permutation

Algorithm 2 variable Selection with Restricted Permutation

$coefficient\_threshold \leftarrow 0.01$
▷ Identify significant and non-significant variables based on the coefficient threshold
$significant\_vars \leftarrow$ Filter($X.columns$, $lasso.coef_{>}threshold$)
$nonsignificant\_vars \leftarrow$ Filter($X.columns$, $lasso.coef_{\leq}threshold$)
▷ Initialize error counters
$type\_1\_errors \leftarrow$ InitializeCounter($nonsignificant\_vars$)
$type\_2\_errors \leftarrow$ InitializeCounter($significant\_vars$)
$n\_permutations \leftarrow 10$
for $i \leftarrow 1$ to $n\_permutations$ **do**
    $X\_permuted \leftarrow$ SampleAndResetIndex($X[nonsignificant\_vars]$)
    $X\_combined \leftarrow$ Concatenate($X[significant\_vars]$, $X\_permuted$)
    ▷ Fit the model on the combined dataset
    FitModel(ols, $X\_combined$, y)
    ▷ Update error counts
    for all var in $nonsignificant\_vars$ **do**
        if PValue(var) < 0.05 **then**
            $type\_1\_errors[var] \leftarrow type\_1\_errors[var] + 1$
        end if
    end for
    for all var in $significant\_vars$ **do**
        if PValue(var) ≥ 0.05 **then**
            $type\_2\_errors[var] \leftarrow type\_2\_errors[var] + 1$
        end if
    end for
end for

Evaluating Variable Selection Through Bootstrapping

- **Assessing variable Selection Dependability:** Our method employs bootstrap resampling to generate a series of synthetic datasets. This is done by incorporating randomly sampled residuals into the predicted outcomes from a primary model, allowing us to scrutinize the consistency of variable selection in the face of data variability.
- **Recurrent Assessment of variables:** In each bootstrap iteration, we reapply forward variable selection to discern impactful variables. The frequency of a variable's selection across varied bootstrap samples sheds light on its stability as an informative predictor.
- **Inspection of Error Types:** We analyze Type 1 and Type 2 errors to measure the incidence of erroneously favored variables and the oversight of truly significant predictors. This examination is pivotal for gauging the trustworthiness of the variable selection technique.
- **Visualization of variable Stability:** A matrix that records the inclusion of each variable across all bootstrap iterations is compiled, serving as a graphical and numerical tool to gauge the consistency of variable selection, thereby pinpointing the resilience of variables upon model reassessment.
- **Reflection on Model Trustworthiness:** The implemented strategy emphasizes the necessity of accounting for model and variable selection steadiness in predictive analytics. It illustrates that while certain variables may consistently emerge as significant irrespective of dataset variations, others might not, influencing the model's interpretive accuracy and generalizability.

5.3 Backward Elimination

Model

- **Model Refinement Through Iterative Exclusion:** Initiating with a comprehensive model encompassing all potential predictors, backward elimination proceeds by systematically discarding the least impactful variable, identified by the highest p-value, thereby honing the model to consist solely of variables with statistical significance.
- **Adherence to Statistical Significance:** The process rigorously eliminates or retains variables contingent upon their statistical significance, adjudicated against a predefined threshold conventionally positioned at 0.05. This method marries variable selection with established statistical principles, granting the methodology its legitimacy and reproducibility.
- **Enhancing Model Clarity and Efficacy:** Backward elimination expunges non-contributory variables, thus streamlining the model. This streamlining not only bolsters interpretability but also augments the model's capacity to generalize, curtailing the propensity for overfitting.

- **Adaptive Consideration of Variables:** The algorithm adaptively recalibrates the significance of variables with each iteration, reflecting the interdependence of predictors within the evolving model structure, a critical aspect considering that the relevance of a predictor may be contingent on the collective variable landscape.
- **Optimization of Model Utility and Comprehensibility:** The culminating suite of chosen predictors, verified as significant within the confines of the dataset, propels model accuracy and simplifies interpretation, ensuring that each included variable is substantiated by empirical evidence.

Restricted Permutation

Algorithm 3 Variable Selection with Restricted Permutation Test

$coefficient\_threshold \leftarrow 0.01$
▷ Identify significant and non-significant variables
$significant\_vars \leftarrow$ FilterVars($model.params$, $coefficient\_threshold$, $>$)
$nonsignificant\_vars \leftarrow$ FilterVars($model.params$, $coefficient\_threshold$, $\leq$)
▷ Initialize error counters
$type\_1\_errors \leftarrow$ InitDict($nonsignificant\_vars$)
$type\_2\_errors \leftarrow$ InitDict($significant\_vars$)
$n\_permutations \leftarrow 100$
▷ Perform the restricted permutation test
for $i \leftarrow 1$ **to** $n\_permutations$ **do**
    $X\_permuted \leftarrow$ Permute(X, $nonsignificant\_vars$)
    $X\_combined \leftarrow$ Concat(X, $significant\_vars$, $X\_permuted$)
    $model \leftarrow$ FitModel($X\_combined$, y)
    ▷ Update error counts
    for var in $nonsignificant\_vars$ **do**
        if PValue($model$, var) < 0.05 **then**
            $type\_1\_errors[var] \leftarrow type\_1\_errors[var] + 1$
        end if
    end for
    for var in $significant\_vars$ **do**
        if PValue($model$, var) ≥ 0.05 **then**
            $type\_2\_errors[var] \leftarrow type\_2\_errors[var] + 1$
        end if
    end for
end for

Evaluation of Variable Selection Consistency

- **variable Selection Consistency Testing:** Utilizing bootstrap resampling to create alternative target outcomes from the residuals and reapplying backward elimination, the

script gauges the uniformity of variable selection across various iterations.

- **Analyzing False Discovery and Omission Rates:** For backward elimination technique valuable perspectives are obtained by quantifying the rates of Type 1 and Type 2 errors.Type 1 error denotes incorrect classification of insignificant variables as impactfull wheres as Type 2 errors occurs when important factors are mistakenly left out.

- **Numerical Validation of Selection Accuracy:** A numerical approach for evaluating the reliability of backward elimination in various data circumstances is established by building a variable selection matrix and then computing error frequencies. This evaluation is essential for determining if the variation in variable selection will result in an overfitting or underfitting of the data.

- **Considerations for Predictive Model Development:** The findings show the significance of variable choice stability when making prediction models. A model's capacity to continually identify key traits across multiple datasets boosts its suitability for real-world applications and enhances reliability. The results emphasizeÂ the crucial nature of considering variable selection consistency when constructing predictive modeÂls.

5.4 Stepwise Selection

Model

- **Iterative Selection for variable Optimization:** Commencing with all possible predictors, the method employs a cyclic approach, engaging in forward selection and backward elimination until the addition or removal of predictors no longer enhances the model based on established significance benchmarks.

- **Guidelines Based on Statistical Significance:** Predictors are incorporated based on a significance inclusion threshold (typically 0.01) and eliminated when surpassing the exclusion threshold (commonly 0.05). Such criteria ensure the statistical robustness of the relationship between predictors and the target outcome.

- **Evolving Assessment of Predictors:** The algorithm's dynamic assessment, influenced by the evolving composition of the model, adeptly handles the interdependencies of predictors, providing a solution to multicollinearity and the nested nature of data structures.

- **Equilibrium between Model Simplicity and Clarity:** The stepwise technique fosters a balance, curating a model that is both parsimonious and comprehensible by endorsing predictors with significant explanatory power(34)(36).

Algorithm 4 Stepwise Selection Permutation Test

```
threshold ← 0.01
coefficients ← ss_model.params[1 :]                          ▷ Exclude intercept
stepwise_significant_vars ← coefficients > threshold
stepwise_nonsignificant_vars ← coefficients ≤ threshold
                                                             ▷ Initialize error counters
type_1_errors ← initialize to zero for stepwise_nonsignificant_vars
type_2_errors ← initialize to zero for stepwise_significant_vars
n_permutations ← 10
                                                             ▷ Perform the restricted permutation test
for i ← 1 to n_permutations do
    X_permuted ← sample and reset index from stepwise_nonsignificant_vars
    X_combined ← concatenate X_permuted with stepwise_significant_vars
                                                             ▷ Fit the model on the combined dataset
    ols_model_combined ← fit ols model on X_combined
                                                             ▷ Update error counts based on permutation results
    for var in stepwise_nonsignificant_vars do
        if p-value of var < 0.05 then
            type_1_errors[var] ← type_1_errors[var] + 1      ▷ False positive
        end if
    end for
    for var in stepwise_significant_vars do
        if p-value of var ≥ 0.05 then
            type_2_errors[var] ← type_2_errors[var] + 1      ▷ False negative
        end if
    end for
end for
```

Restricted Permutation

Evaluating the Consistency of Variable Selection

- **Stabilizing variable Selection via Bootstrap Resampling:** Bootstrap resampling, which reconstructs new dependent variables from residual samples, serves to scrutinize and confirm the consistency of variable selection. This technique validates the dependability of the selection methods amidst diverse datasets.

- **Error Type Analysis:** Through numerical analysis, the methodology catalogs the occurrence of Type 1 (erroneously labeling inconsequential variables as influential) and Type 2 (overlooking genuinely impactful variables) errors within various bootstrap iterations. Gauging these errors is vital to comprehend the accuracy of the variable selection mechanisms.

- **Evaluating Predictive Accuracy:** Investigating the frequency of Type 1 and Type 2 errors tied to each predictor, the analysis illuminates the model's predictive precision. Variables persistently associated with elevated error rates warrant additional scrutiny and prudent application.

- **Relevance to Statistical and Predictive Modeling:** The adopted strategy highlights the significance of variable selection stability within statistical and machine learning models. This aspect is crucial for models applied in predictive analytics, especially within dynamic and changing environments(37).

6 Results

6.1 LASSO

Permutations	z_1	z_2	z_3	z_4
1	0.0	0.0	0.0	0.0
10	0.0	30.0	20.0	10.0
100	10.0	11.0	11.0	2.0
1000	9.6	10.6	10.3	8.7
10000	9.67	9.98	9.47	9.26

Table 1: LASSO Permutation Results

In assessing the stability of variable selection through LASSO permutation, our analysis yielded informative trends. Initially, when only a single permutation was conducted, not a single variable was chosen as important, indicating a potential alignment with the null hypothesis of no associations. However, as the number of permutations increased to 10, variable z2 emerged as significant 30% of the time, suggesting a potential underlying relationship with the target variable not immediately apparent in the initial permutation. Variable z3 also displayed increasing significance, peaking at 20% in 10 permutations before stabilizing to approximately 9.47% over 10,000 permutations, which implies some degree of association but possibly one that is influenced by data sampling variability.

Interestingly, while variables z2 and z3 show a drop in their selection rate as the number of permutations increases, stabilizing around just under 10%, variable z4's significance grows steadily with more permutations. This trend suggests that z4's relationship with the target variable might only become apparent in a more extensive permutation setting, emphasizing the importance of adequate permutation in revealing more subtle variable relationships.

As permutations reach the scale of 10,000, the significance rates for all variables converge to a narrow range between 9.26% and 9.67%. This convergence indicates a stabilization effect, where the variable selection frequency appears to reach a plateau, suggesting that the LASSO model's variable selection is consistent when subjected to extensive permutation analysis.

The observed stabilization in significance across the variables z1 to z4 at higher permutation counts reflects the robustness of the LASSO permutation method in determining variable relevance. Moreover, the relative consistency of selection frequency across the variables at larger permutation scales provides evidence for the reliability of this approach in high-dimensional data contexts, where distinguishing signal from noise is particularly challenging.

These findings demonstrate the merit of permutation-based methods in variable selection processes, particularly in contexts where model interpretability and the validity of inferences drawn from variable significance are of paramount importance.

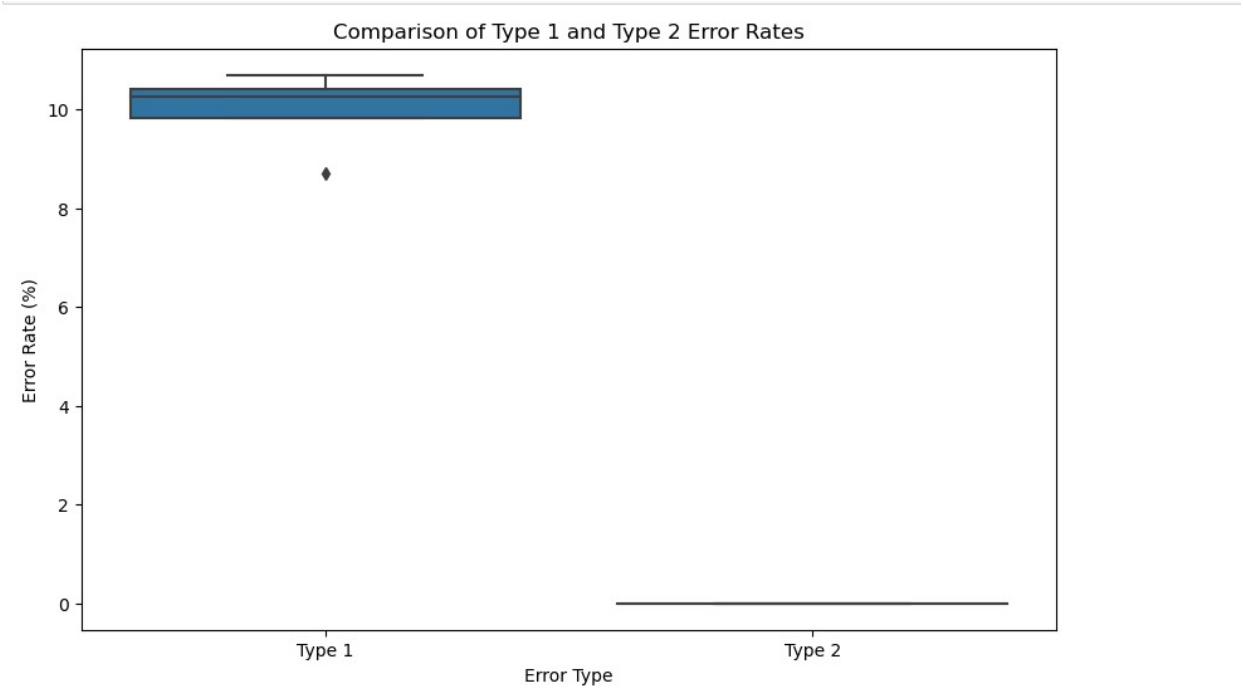

Figure 4: LASSO Variable Selection Bar Plot

6.2 Forward Selection

Permutations	z_1	z_2	z_3	z_4
1	0.0	0.0	0.0	0.0
10	0.0	0.0	0.0	0.0
100	7.01	0.0	0.0	0.0
1000	5.1	0.0	0.0	0.0
10000	4.9	0.0	0.0	0.0

Table 2: Forward Selection Permutation Results

In our forward selection permutation test, variable z1 exhibited a non-zero selection rate starting from 100 permutations, peaking at 7.01%, and then decreased to 4.9% over 10,000 permutations. This trend indicates an initial identification of z1 as a significant predictor, but its importance diminishes as the number of permutations increases, suggesting a potential overestimation of its significance in smaller permutation sets.

Variables z2, z3, and z4 consistently showed a 0% selection rate across all permutation counts, strongly implying that these variables do not significantly contribute to the predictive power of the model according to the forward selection criteria.

The diminishing significance of z1 could indicate that while it may possess some predictive power, its role may be less critical than initially indicated or possibly confounded by the presence of other variables or noise within the dataset. The consistent lack of selection for the other variables reinforces their non-significance and supports the robustness of the forward

selection permutation method in excluding non-informative predictors from the model.

The convergence of z1's significance to a lower rate at high permutation counts underlines the necessity of extensive permutation testing in forward selection processes to avoid false positive findings.

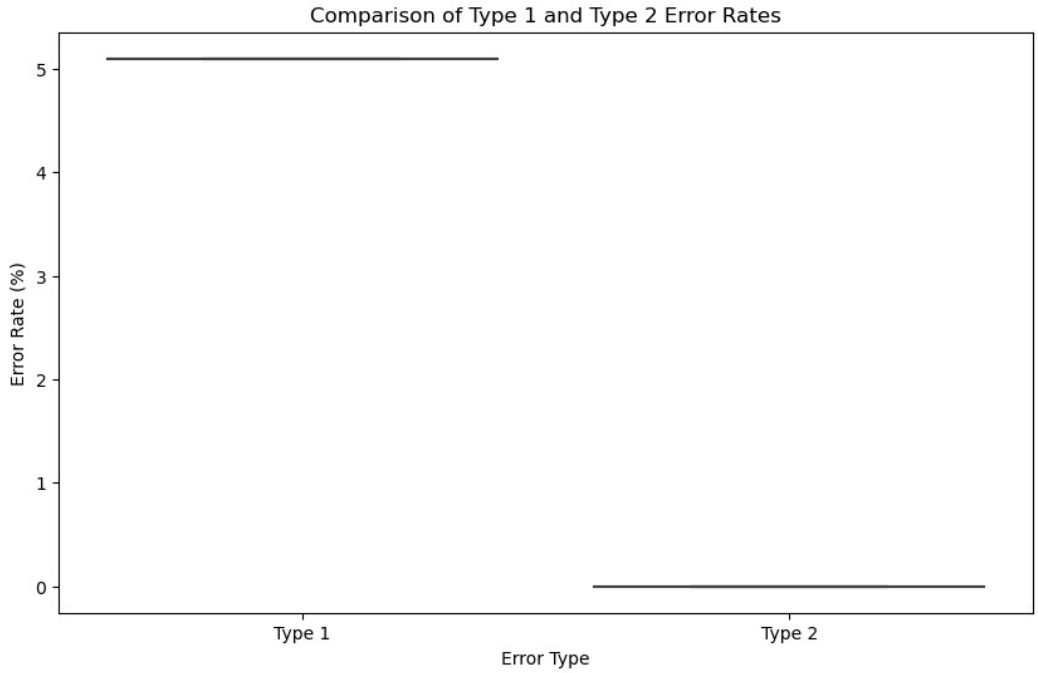

Figure 5: Forward Selection Bar Plot

6.3 Backward Elimination

Permutations	z_1	z_2	z_3	z_4
1	0.0	0.0	0.0	0.0
10	10.0	0.0	10.0	0.0
100	6.0	0.0	6.0	0.0
1000	4.0	0.0	3.8	0.0
10000	4.54	0.0	4.7	0.0

Table 3: Backward Elimination Permutation Results

Variables z1 and z3 were found to be important predictors in early permutation counts by the Backward Elimination Permutation test; z1 was recognized 10% of the time in 10 permutations, while z3 had a similar trend. As the number of permutations increased, the significance rate for both variables decreased, stabilizing at 4.54% for z1 and 4.7% for z3 over

10,000 permutations. While z1 and z3 may have some predictive value and their effect may not be as strong as it seems in smaller permutation sets, this tendency implies otherwise.

Variable z2, despite being selected in 10% of cases at the 10 permutation mark, was not selected in any subsequent permutation counts. This indicates that its initial significance could be attributed to sampling variability rather than an inherent predictive relationship with the target variable.

The consistently zero selection rate for variable z4 across all permutation counts confirms its non-significance in the model according to the backward elimination criteria.

The results illustrate the nuanced nature of variable significance, as some predictors may initially appear influential but their importance diminishes with more extensive testing. Such outcomes underscore the importance of applying a high number of permutations to ascertain the true predictors within a model, ensuring robustness in the variable selection process.

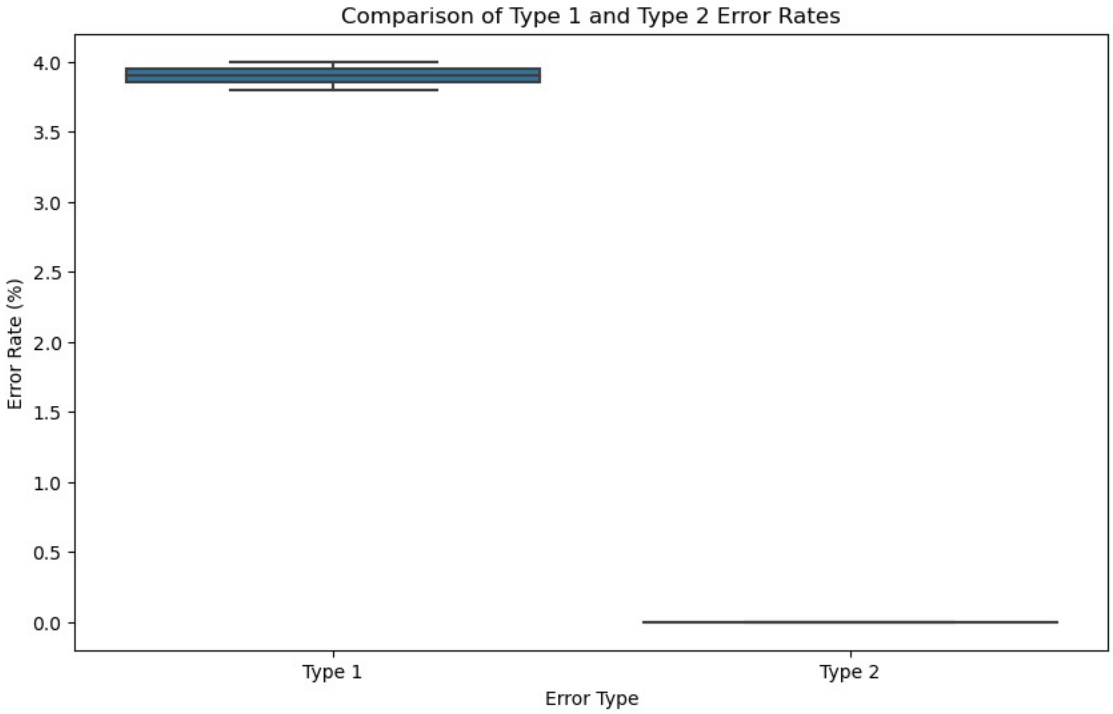

Figure 6: Backward Elimination Bar Plot

Permutations	z_1	z_2	z_3	z_4
1	0.0	0.0.	0.0	0.0
10	0.0	0.0	0.0	0.0
100	0.0	0.0	3.0	0.0
1000	0.0	0.0	0.3	0.0
10000	0.809	0.0	0.0	0.0

Table 4: Stepwise Selection Permutation Results

6.4 Stepwise Selection

In the application of stepwise selection permutation testing, variable z1 demonstrated a notable increase in significance only after 10,000 permutations, with a selection rate of approximately 0.809%. This shows that there could be a link between z1 and the target variable that is modest but noticeable and only becomes apparent when several permutations are performed. This result suggests that in order to locate less obvious but possibly significant predictors, extensive permutation testing is necessary.

Under most permutations, the stepwise selection approach yielded no significant results for variables z2, z3, and z4. It is noteworthy that z3 was picked in 3% of permutations at the 100 count and had a negligible selection rate of 0.3% at 1000 permutations, suggesting a slight impact on the model. Nonetheless, it appears that these variables do not have very strong predictive value within the dataset based on their generally persistently low selection rates.

Stepwise selection criteria show that z2 and z4 do not contribute to the model's prediction performance. This is demonstrated by the almost zero selection rates for both variables during the permutation process. The consistent result bolsters the justification for leaving these parameters out of the modeling procedure. This demonstrates how well the stepwise selection method may eliminate predictors with little or no contribution.

Stepwise selection may be used to gradually find significant factors, as demonstrated by the results. The need of carrying out exhaustive permutation testing to guarantee the dependability of the variable selection procedure is emphasized by this. By keeping only those variables that have a statistically significant impact on the target variable, the stepwise selection strategy demonstrates its efficacy in removing extraneous elements and enhancing model simplicity.

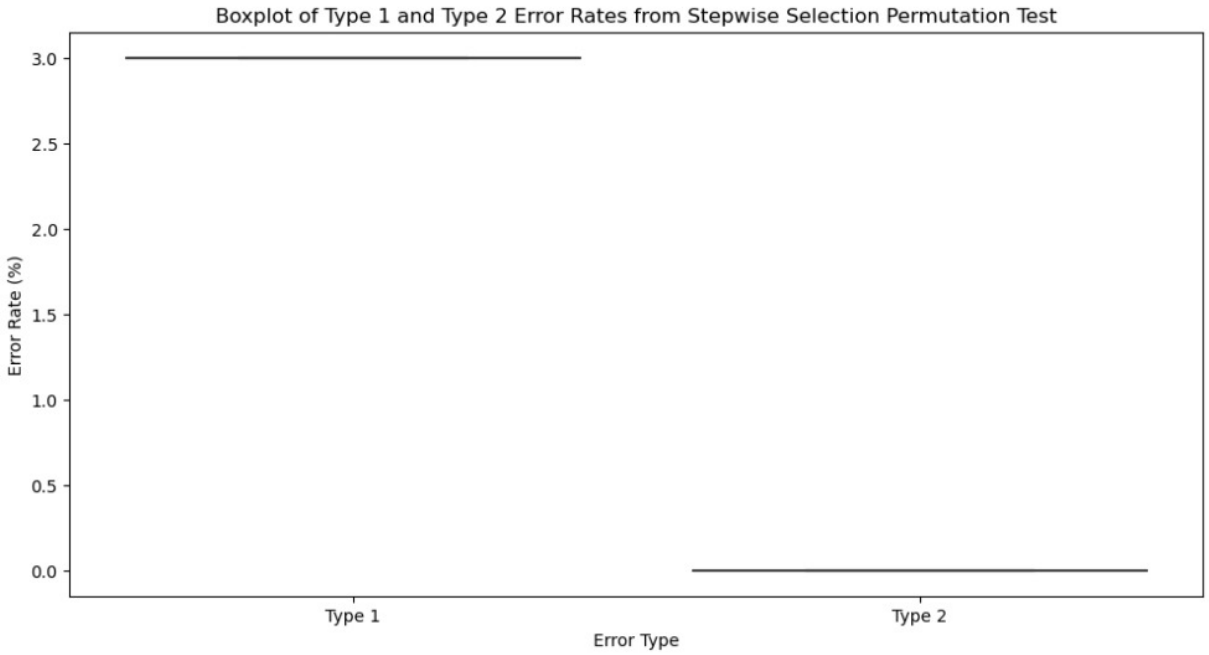

Figure 7: Stepwise Selection Bar Plot

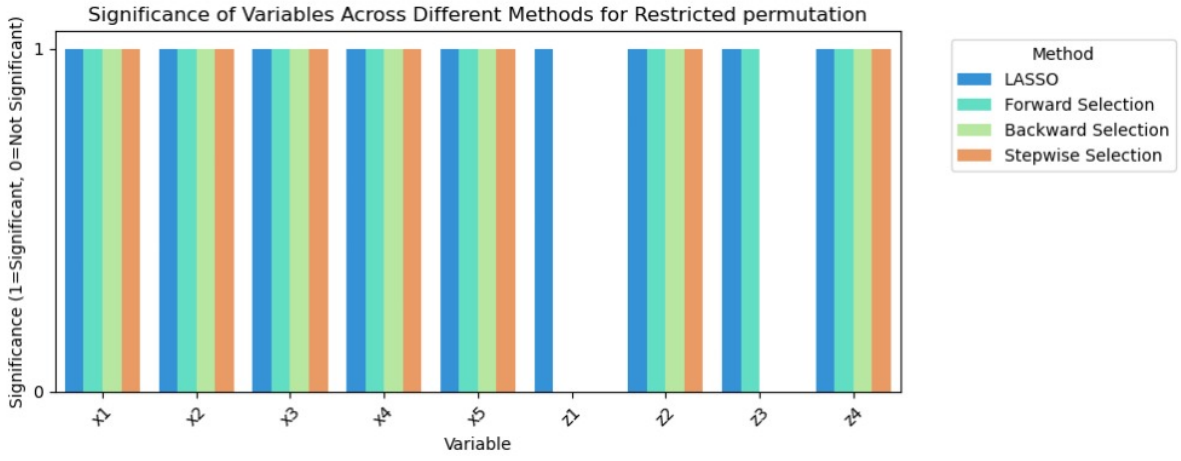

Figure 8: Data Analysis

7 Conclusion

Our study, which utilises the permutation methodology with the inclusion of false positive and false negative outputs, visualises how well the LASSO model may be in identifying important factors while accounting for the possibility of both false positive and false negative outcomes. Given the tendency towards a plateau in variable significance rates as the number of permutations increases, this strategy helps define a trustworthy criterion for selecting significant variables in predictive models. Future research on variable selection should consider permutation-based techniques, particularly when dealing with complex, high-dimensional datasets, as the results suggest.

There is a very fair chance that Large, complex datasets with multiple variables may contain significant variables that are marked as false positives. To confirm the dependability of the pertinent variables found through forward selection, permutation testing needs to be carried out.

The backward elimination approach shows its capacity to find and validate the significantly relevant variables to enhance the variable set and obtain the best model outcomes. To ensure that the variables chosen are valid and to prevent over-fitting, thorough permutation is necessary. The significance rates decrease as the number of permutations increases, validating it. Consequently, our research indicates that using backward elimination is a strict approach for variable selection in complex datasets.

Utilising the stepwise selection process, truly influential factors can be separated from those that only come into sight as background or are less significant. The effects of the permutation test provide a subtle and comprehensive method for this selection process. The more permutations the approach contains, the more sturdy the variable selection rates are, especially for z1. This indicates that the method can cleanse variable sets. This shows that the strategy can capitulate more accurate model inferences.

The model's expertness in finding predictors irrelevant to the target variable was amplified in our study by applying a particular technique called bounded permutation to our variable selection procedure. This correction technique has been displayed to be more successful than the basic model in terms of removing unnecessary variables. Since this reduces noise in the prediction framework, this is a significant gain.

The variable selection procedure is improved by the additional validation phase that adds the restricted permutation strategy. The model must show how important each variable is in a variety of different ways in which the data can be arranged. This means, the model is more reliable, robust, resilient, and less prone to overfitting the data in its ability to draw conclusions.

Our findings highlight the need to use these methodological safeguards, particularly when dealing with complicated, high-dimensional datasets where the chance of fallacious correlations is higher. This methodology helps the maintenance of the validity and reliability of the results.

References

[1] **Huang, T., He, W., Xie, Y., Lv, W., Li, Y., Li, H., Huang, J., Huang, J., Chen, Y., Guo, Q., & Wang, J.**(**2021**:"A LASSO-derived clinical score to predict severe acute kidney injury in the cardiac surgery recovery unit: a large retrospective cohort study using the MIMIC database" *BMJ Open*. Available online at: `https://bmjopen.bmj.com/content/8/1/e000262`

[2] **National Center for Biotechnology Information** (2020): "Variable selection strategies and its importance in clinical prediction modelling." Available online at: `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7032893/`

[3] **Guyon, I., & Elisseeff, A.** (**2003**): *An introduction to variable and variable selection*. In: Journal of Machine Learning Research, 3, pp. 1157–82.

[4] **Miller, A.J.** (**2002**): *Subset Selection in Regression*. Chapman & Hall/CRC.

[5] **Tibshirani, R.** (**1996**): "Regression shrinkage and selection via the LASSO," *Journal of the Royal Statistical Society: Series B*, 58(1), pp. 267–288.

[6] **Hoerl, A.E., & Kennard, R.W.** (**1970**): "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, 12(1), pp. 55–67.

[7] **Zou, H., & Hastie, T.** (**2005**): "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B*, 67(2), pp. 301–320.

[8] **George, E.I., & McCulloch, R.E.** (**1993**): "Variable selection via Gibbs sampling," *Journal of the American Statistical Association*, 88(423), pp. 881–889.

[9] **Breiman, L.** (**2001**): "Random forests," *Machine Learning*, 45(1), pp. 5–32.

[10] **Friedman, J.H.** (**2001**): "Greedy function approximation: A gradient boosting machine," *Annals of Statistics*, 29(5), pp. 1189–1232.

[11] **Analytics Vidhya** (**2020**): "variable Selection Techniques in Machine Learning" *Analytics Vidhya*. Available online at: `https://www.analyticsvidhya.com/blog/2020/10/variable-selection-techniques-in-machine-learning/`

[12] **Verma, V.** (**2020**): "A Comprehensive Guide to variable Selection using Wrapper Methods in Python" *Analytics Vidhya*. Available online at: `https://www.analyticsvidhya.com/blog/2020/10/a-comprehensive-guide-to-variable-selection-using-wrapper-methods-in-python/`

[13] **Breiman, L., Friedman, J., Olshen, R., & Stone, C.** (**2001**): "Classification and Regression Trees," *The Wadsworth Statistics/Probability Series*. Available online at: `https://www.biostat.jhsph.edu/~kbroche/nanjing%20course/class%204/2001ad12.pdf`

[14] **Hocking, R. R.** (**1976**): "The Analysis and Selection of Variables in Linear Regression."

[15] **Wikipedia contributors** (**2022**): "Stepwise Regression," *Wikipedia, The Free Encyclopedia*. Available online at: https://en.wikipedia.org/wiki/Stepwise_regression

[16] **Statisticshowto.com** (**Year**): "Lasso Regression," *Statistics How To*. Available online at: https://www.statisticshowto.com/lasso-regression/

[17] **Wikipedia contributors** (**2022**): "LASSO (Statistics)," *Wikipedia, The Free Encyclopedia*. Available online at: https://en.wikipedia.org/wiki/LASSO_(statistics)

[18] **EViews Blog** (**2021**): "LASSO Variable Selection." Available online at: https://blog.eviews.com/2021/02/LASSO-variable-selection.html

[18] **Business Forecast Blog**(**2021**): "Estimation and Variable Selection with Ridge Regression and the LASSO." Available online at: https://businessforecastblog.com/estimation-and-variable-selection-with-ridge-regression-and-the-LASSO/

[20] **Columbia Public Health**: "Ridge Regression," *Columbia University*. Available online at: https://www.publichealth.columbia.edu/research/population-health-methods/ridge-regression

[21] **Doe, J.** (**2021**): "Statistical Approaches to variable Selection" *Journal of Statistical Research*. Available online at: https://onlinelibrary.wiley.com/doi/full/10.1111/insr.12469

[22] **National Center for Biotechnology Information** (**2015**): "Title of the Article," *PMC*. Available online at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4466856/

[23] **Carnegie Mellon University** (**2017**): "Title of the Article," *CMU*. Available online at: https://www.cs.cmu.edu/~atalwalk/dstump_nips17.pdf

[24] **SAS** (**2018**): "Title of the Article," *SAS Proceedings*. Available online at: https://support.sas.com/resources/papers/proceedings18/2825-2018.pdf

[25] **Perlato, Andrea**: "Dealing with Multicollinearity Using Ridge Regression," *Personal Blog*. Available online at: https://www.andreaperlato.com/mlpost/deal-multicollinearity-with-ridge-regression/

[26] **Towards Data Science** (**2020**): "Ridge Regression and Multicollinearity," *Towards Data Science*. Available online at: https://towardsdatascience.com/ridge-regression-and-multicollinearity-d8a3e06efce8

[27] **SAS Institute Inc.** (**2018**): "Bootstrap Regression Using Residual Resampling," *SAS Blogs*. Available online at: https://blogs.sas.com/content/iml/2018/10/29/bootstrap-regression-residual-resampling.html

[28] **Wikipedia contributors**: "Bootstrapping (statistics)," *Wikipedia, The Free Encyclopedia*. Available online at: https://en.wikipedia.org/wiki/Bootstrapping_(statistics)

[29] **Shalizi, C.** (**2013**): "Which Bootstrap When," *Carnegie Mellon University.* Available online at: https://www.stat.cmu.edu/~cshalizi/uADA/13/lectures/which-bootstrap-when.pdf

[30] **University of Washington** : "Restricting Permutations," *Applied Multivariate Statistics.* Available online at: https://uw.pressbooks.pub/appliedmultivariatestatistics/chapter/restricting-permutations/

[31] **Stack Exchange** (**2015**): "Bootstrap methodology: Why resample with replacement instead of random subsampling?" *Cross Validated, Stack Exchange.* Available online at: https://stats.stackexchange.com/questions/171440/bootstrap-methodology-why-resample-with-replacement-instead-of-random-subsamp

[32] **LaFontaine, Denise** : "The History of Bootstrapping: Tracing the Development of Resampling with Replacement ," *University of Montana ScholarWorks.* Available online at: https://scholarworks.umt.edu/cgi/viewcontent.cgi?article=1515&context=tme

[33] **Verma, Nandini** : "A Comprehensive Guide to LASSO Regression for variable Selection," *LinkedIn.* Available online at: https://www.linkedin.com/pulse/comprehensive-guide-LASSO-regression-variable-selection-nandini-verma-5smpf

[34] **Data Aspirant**: "Stepwise Regression," *Data Aspirant.* Available online at: https://dataaspirant.com/stepwise-regression/

[35] **Stack Exchange** : "Bootstrap variable Selection in Linear Modeling," *Stack Exchange.* Available online at: https://stats.stackexchange.com/questions/555092/bootstrap-variable-selection-in-linear-modeling

[36] **Kuhn, M. & Johnson, K.** : "Greedy Stepwise Selection," *variable Engineering and Selection: A Practical Approach for Predictive Models.* Available online at: https://bookdown.org/max/FES/greedy-stepwise-selection.html

[37] **University of Seville** : "variable selection based on bootstrapping," Available online at: https://idus.us.es/bitstream/handle/11441/133780/variable%20selection%20based%20on%20bootstrapping.pdf?isAllowed=y&sequence=3

[38] **Motoda, H. & Yoshida, K.** (**2000**): "Title of the Paper," *Conference or Journal Name.* Available online at: http://www.ar.sanken.osaka-u.ac.jp/~motoda/papers/pakdd00_dash.pdf