

Data Mining & Machine Learning

Regularization Methods for Regression

Francisco Javier Arceo
Senior Data Scientist

NYU Polytechnic School of Engineering
Commonwealth Bank of Australia

March 23, 2015

L2 Norm: Ridge

Motivation

Framework

Similarities

L1 Norm: Lasso

Similarities

Motivation

Framework

Regularization Methods

Important Facts

Empirical Example: Kaggle's Amazon Competition

Amazon Data: LASSO Model

Amazon Data: Ridge Model

Amazon Data: Performance

Some Resources

Ridge Regression

$$\hat{\beta}_{ridge} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^n (y_i - \sum_{j=1}^p \beta_j x_{ij})^2 + \lambda \sum_{j=1}^p \beta_j^2 \quad (1)$$

$$\hat{\beta}_{ridge} = \underset{\beta}{\operatorname{argmin}} (\mathbf{y} - \mathbf{x}^T \beta)^2 + \lambda \beta^2 \quad (2)$$

- First published by the Russian Mathematician Andrey Nikolayevich Tikhonov (1906-1993) in 1943

Ridge Regression

$$\hat{\beta}_{ridge} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^n (y_i - \sum_{j=1}^p \beta_j x_{ij})^2 + \lambda \sum_{j=1}^p \beta_j^2 \quad (1)$$

$$\hat{\beta}_{ridge} = \underset{\beta}{\operatorname{argmin}} (\mathbf{y} - \mathbf{x}^T \beta)^2 + \lambda \beta^2 \quad (2)$$

- ▶ First published by the Russian Mathematician Andrey Nikolayevich Tikhonov (1906-1993) in 1943
- ▶ Originally called Tikhonov regularization

Ridge Regression

$$\hat{\beta}_{ridge} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^n (y_i - \sum_{j=1}^p \beta_j x_{ij})^2 + \lambda \sum_{j=1}^p \beta_j^2 \quad (1)$$

$$\hat{\beta}_{ridge} = \underset{\beta}{\operatorname{argmin}} (\mathbf{y} - \mathbf{x}^T \beta)^2 + \lambda \beta^2 \quad (2)$$

- ▶ First published by the Russian Mathematician Andrey Nikolayevich Tikhonov (1906-1993) in 1943
- ▶ Originally called Tikhonov regularization
- ▶ First introduced to the statistical literature in 1970 by Hoerl and Kennard (*Technometrics*)

Ridge Regression

$$\hat{\beta}_{ridge} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^n (y_i - \sum_{j=1}^p \beta_j x_{ij})^2 + \lambda \sum_{j=1}^p \beta_j^2 \quad (1)$$

$$\hat{\beta}_{ridge} = \underset{\beta}{\operatorname{argmin}} (\mathbf{y} - \mathbf{x}^T \beta)^2 + \lambda \beta^2 \quad (2)$$

- ▶ First published by the Russian Mathematician Andrey Nikolayevich Tikhonov (1906-1993) in 1943
- ▶ Originally called Tikhonov regularization
- ▶ First introduced to the statistical literature in 1970 by Hoerl and Kennard (*Technometrics*)
- ▶ Method to deal with non-invertible $(\mathbf{X}'\mathbf{X})$ matrices

Ridge Regression

$$\hat{\beta}_{j,ridge} = \frac{\sum_{i=1}^n (x_{ij} - \bar{x}_j)(y_i - \bar{y})}{\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2} \frac{1}{1 + \lambda} = \frac{\hat{\beta}_{j,ols}}{1 + \lambda} \quad (3)$$

$$\hat{\beta}_{ridge} = (\mathbf{X}'\mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}'\mathbf{y} \quad (4)$$

- Equation (3) is a special case of an orthonormal design matrix

Ridge Regression

$$\hat{\beta}_{j,ridge} = \frac{\sum_{i=1}^n (x_{ij} - \bar{x}_j)(y_i - \bar{y})}{\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2} \frac{1}{1 + \lambda} = \frac{\hat{\beta}_{j,ols}}{1 + \lambda} \quad (3)$$

$$\hat{\beta}_{ridge} = (\mathbf{X}'\mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}'\mathbf{y} \quad (4)$$

- ▶ Equation (3) is a special case of an orthonormal design matrix
- ▶ λ is a penalty term that is added to the diagonal elements of $(\mathbf{X}'\mathbf{X})$ and \mathbf{I} is an $(n \times n)$ matrix

Ridge Regression

$$\hat{\beta}_{j,ridge} = \frac{\sum_{i=1}^n (x_{ij} - \bar{x}_j)(y_i - \bar{y})}{\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2} \frac{1}{1 + \lambda} = \frac{\hat{\beta}_{j,ols}}{1 + \lambda} \quad (3)$$

$$\hat{\beta}_{ridge} = (\mathbf{X}'\mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}'\mathbf{y} \quad (4)$$

- ▶ Equation (3) is a special case of an orthonormal design matrix
- ▶ λ is a penalty term that is added to the diagonal elements of $(\mathbf{X}'\mathbf{X})$ and \mathbf{I} is an $(n \times n)$ matrix
- ▶ Question: what happens when $\lambda=0$?

Ridge Regression

$$\hat{\beta}_{j,ridge} = \frac{\sum_{i=1}^n (x_{ij} - \bar{x}_j)(y_i - \bar{y})}{\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2} \frac{1}{1 + \lambda} = \frac{\hat{\beta}_{j,ols}}{1 + \lambda} \quad (3)$$

$$\hat{\beta}_{ridge} = (\mathbf{X}'\mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}'\mathbf{y} \quad (4)$$

- ▶ Equation (3) is a special case of an orthonormal design matrix
- ▶ λ is a penalty term that is added to the diagonal elements of $(\mathbf{X}'\mathbf{X})$ and \mathbf{I} is an $(n \times n)$ matrix
- ▶ Question: what happens when $\lambda=0$?
- ▶ Back to Ordinary Least-Squares solution!

Ridge Regression

$$\hat{\beta}_{ridge} = \underset{\beta}{\operatorname{argmin}} (\mathbf{y} - \mathbf{x}^T \beta)^2 + \lambda \beta^2 \quad (5)$$

$$\hat{\beta}_{ridge} = \underset{\beta}{\operatorname{argmin}} \left(\sum_{i=1}^n (y_i - x_i^T \beta)^2 \right) \text{ s.t. } \|\beta\|_2 \leq t \quad (6)$$

- Equation (5) is equivalent to solving the constrained optimization problem in equation (6)

Ridge Regression

$$\hat{\beta}_{ridge} = \underset{\beta}{\operatorname{argmin}} (\mathbf{y} - \mathbf{x}^T \beta)^2 + \lambda \beta^2 \quad (5)$$

$$\hat{\beta}_{ridge} = \underset{\beta}{\operatorname{argmin}} \left(\sum_{i=1}^n (y_i - x_i^T \beta)^2 \right) \text{ s.t. } \|\beta\|_2 \leq t \quad (6)$$

- ▶ Equation (5) is equivalent to solving the constrained optimization problem in equation (6)
- ▶ Note that the $\|\beta\|_2$ is the 2-norm of the vector β and $t \geq 0$

Ridge Regression

$$\hat{\beta}_{ridge} = \underset{\beta}{\operatorname{argmin}} (\mathbf{y} - \mathbf{x}^T \beta)^2 + \lambda \beta^2 \quad (5)$$

$$\hat{\beta}_{ridge} = \underset{\beta}{\operatorname{argmin}} \left(\sum_{i=1}^n (y_i - x_i^T \beta)^2 \right) \text{ s.t. } \|\beta\|_2 \leq t \quad (6)$$

- ▶ Equation (5) is equivalent to solving the constrained optimization problem in equation (6)
- ▶ Note that the $\|\beta\|_2$ is the 2-norm of the vector β and $t \geq 0$
- ▶ What happens when we change the norm? What about the L1 norm?

LASSO Regression

$$\hat{\beta}_{ridge} = \underset{\beta}{\operatorname{argmin}} \left(\sum_{i=1}^n (y_i - x_i^T \beta)^2 \right) \text{ s.t. } \|\beta\|_2 \leq t \quad (7)$$

$$\hat{\beta}_{lasso} = \underset{\beta}{\operatorname{argmin}} \left(\sum_{i=1}^n (y_i - x_i^T \beta)^2 \right) \text{ s.t. } \|\beta\|_1 \leq t \quad (8)$$

- ▶ $\|\beta\|_1 = \sum_{j=1}^p |\beta_j|$, $\|\beta\|_2 = \sqrt{\sum_{j=1}^p \beta_j^2}$, and $p :=$ the number of features

LASSO Regression

$$\hat{\beta}_{ridge} = \underset{\beta}{\operatorname{argmin}} \left(\sum_{i=1}^n (y_i - x_i^T \beta)^2 \right) \text{ s.t. } \|\beta\|_2 \leq t \quad (7)$$

$$\hat{\beta}_{lasso} = \underset{\beta}{\operatorname{argmin}} \left(\sum_{i=1}^n (y_i - x_i^T \beta)^2 \right) \text{ s.t. } \|\beta\|_1 \leq t \quad (8)$$

- ▶ $\|\beta\|_1 = \sum_{j=1}^p |\beta_j|$, $\|\beta\|_2 = \sqrt{\sum_{j=1}^p \beta_j^2}$, and $p :=$ the number of features
- ▶ $\|\beta\|_1$ is the unit-norm of the vector β and $t \geq 0$

LASSO Regression

$$\hat{\beta}_{ridge} = \underset{\beta}{\operatorname{argmin}} \left(\sum_{i=1}^n (y_i - x_i^T \beta)^2 \right) \text{ s.t. } \|\beta\|_2 \leq t \quad (7)$$

$$\hat{\beta}_{lasso} = \underset{\beta}{\operatorname{argmin}} \left(\sum_{i=1}^n (y_i - x_i^T \beta)^2 \right) \text{ s.t. } \|\beta\|_1 \leq t \quad (8)$$

- ▶ $\|\beta\|_1 = \sum_{j=1}^p |\beta_j|$, $\|\beta\|_2 = \sqrt{\sum_{j=1}^p \beta_j^2}$, and $p :=$ the number of features
- ▶ $\|\beta\|_1$ is the unit-norm of the vector β and $t \geq 0$
- ▶ Difference is in the constraint of the minimization problem

LASSO Regression

$$\hat{\beta}_{lasso} = \underset{\beta}{\operatorname{argmin}} \left(\sum_{i=1}^n (y_i - x_i^T \beta)^2 \right) \text{ s.t. } \|\beta\|_1 \leq t \quad (9)$$

$$\hat{\beta}_{lasso} = \underset{\beta}{\operatorname{argmin}} (\mathbf{y} - \mathbf{x}^T \beta)^2 + \lambda |\beta| \quad (10)$$

- Equation (9) and (10) lead to a minimization similar to equation (2) from Ridge, except the penalization to β is linear instead of quadratic

LASSO Regression

$$\hat{\beta}_{lasso} = \underset{\beta}{\operatorname{argmin}} \left(\sum_{i=1}^n (y_i - x_i^T \beta)^2 \right) \text{ s.t. } \|\beta\|_1 \leq t \quad (9)$$

$$\hat{\beta}_{lasso} = \underset{\beta}{\operatorname{argmin}} (\mathbf{y} - \mathbf{x}^T \beta)^2 + \lambda |\beta| \quad (10)$$

- ▶ Equation (9) and (10) lead to a minimization similar to equation (2) from Ridge, except the penalization to β is linear instead of quadratic
- ▶ What's the benefit of using the L1-norm versus the L2-norm?

LASSO Regression

$$\hat{\beta}_{lasso} = \underset{\beta}{\operatorname{argmin}} \left(\sum_{i=1}^n (y_i - x_i^T \beta)^2 \right) \text{ s.t. } \|\beta\|_1 \leq t \quad (9)$$

$$\hat{\beta}_{lasso} = \underset{\beta}{\operatorname{argmin}} (\mathbf{y} - \mathbf{x}^T \beta)^2 + \lambda |\beta| \quad (10)$$

- ▶ Equation (9) and (10) lead to a minimization similar to equation (2) from Ridge, except the penalization to β is linear instead of quadratic
- ▶ What's the benefit of using the L1-norm versus the L2-norm?
- ▶ Sparsity!

LASSO Regression

$$\hat{\beta}_{lasso} = \underset{\beta}{\operatorname{argmin}} \left(\sum_{i=1}^n (y_i - x_i^T \beta)^2 \right) \text{ s.t. } \|\beta\|_1 \leq t \quad (9)$$

$$\hat{\beta}_{lasso} = \underset{\beta}{\operatorname{argmin}} (\mathbf{y} - \mathbf{x}^T \beta)^2 + \lambda |\beta| \quad (10)$$

- ▶ Equation (9) and (10) lead to a minimization similar to equation (2) from Ridge, except the penalization to β is linear instead of quadratic
- ▶ What's the benefit of using the L1-norm versus the L2-norm?
- ▶ Sparsity!
- ▶ Using LASSO results in models with less parameters (features)

LASSO Regression

$$\hat{\beta}_{lasso}(\lambda) = S(\hat{\beta}, \lambda) \equiv \text{sign}(\hat{\beta})(|\hat{\beta}| - \lambda)_+ \quad (11)$$

$$S(\hat{\beta}, \lambda) = \begin{cases} \hat{\beta} - \lambda, & \text{if } \hat{\beta} > 0 \text{ and } \lambda < |\hat{\beta}| \\ \hat{\beta} + \lambda, & \text{if } \hat{\beta} < 0 \text{ and } \lambda < |\hat{\beta}| \\ 0, & \text{if } \lambda > |\hat{\beta}|. \end{cases}$$

$$\hat{\beta}_{j,lasso}^t \leftarrow S(\hat{\beta}_{j,lasso}^{t-1} + \sum_{i=1}^n x_{ij}(y_i - \hat{y}_i^{t-1}), \lambda) \quad (12)$$

- Equation (11) is called the soft-threshold operator

LASSO Regression

$$\hat{\beta}_{lasso}(\lambda) = S(\hat{\beta}, \lambda) \equiv \text{sign}(\hat{\beta})(|\hat{\beta}| - \lambda)_+ \quad (11)$$

$$S(\hat{\beta}, \lambda) = \begin{cases} \hat{\beta} - \lambda, & \text{if } \hat{\beta} > 0 \text{ and } \lambda < |\hat{\beta}| \\ \hat{\beta} + \lambda, & \text{if } \hat{\beta} < 0 \text{ and } \lambda < |\hat{\beta}| \\ 0, & \text{if } \lambda > |\hat{\beta}|. \end{cases}$$

$$\hat{\beta}_{j,lasso}^t \leftarrow S(\hat{\beta}_{j,lasso}^{t-1} + \sum_{i=1}^n x_{ij}(y_i - \hat{y}_i^{t-1}), \lambda) \quad (12)$$

- ▶ Equation (11) is called the soft-threshold operator
- ▶ Apply soft-threshold operator as we cycle through our features at the t^{th} iteration (12) updating the residual along the way

LASSO Regression

$$\hat{\beta}_{lasso}(\lambda) = S(\hat{\beta}, \lambda) \equiv \text{sign}(\hat{\beta})(|\hat{\beta}| - \lambda)_+ \quad (11)$$

$$S(\hat{\beta}, \lambda) = \begin{cases} \hat{\beta} - \lambda, & \text{if } \hat{\beta} > 0 \text{ and } \lambda < |\hat{\beta}| \\ \hat{\beta} + \lambda, & \text{if } \hat{\beta} < 0 \text{ and } \lambda < |\hat{\beta}| \\ 0, & \text{if } \lambda > |\hat{\beta}|. \end{cases}$$

$$\hat{\beta}_{j,lasso}^t \leftarrow S(\hat{\beta}_{j,lasso}^{t-1} + \sum_{i=1}^n x_{ij}(y_i - \hat{y}_i^{t-1}), \lambda) \quad (12)$$

- ▶ Equation (11) is called the soft-threshold operator
- ▶ Apply soft-threshold operator as we cycle through our features at the t^{th} iteration (12) updating the residual along the way
- ▶ LASSO does variable selection by setting select variables to 0!

LASSO and Ridge Regression

- ▶ How do we choose λ ? **Cross Validation**

LASSO and Ridge Regression

- ▶ How do we choose λ ? **Cross Validation**
- ▶ When do we use Ridge vs LASSO? When we want a model with less coefficients (predictors) use LASSO

LASSO and Ridge Regression

- ▶ How do we choose λ ? **Cross Validation**
- ▶ When do we use Ridge vs LASSO? When we want a model with less coefficients (predictors) use LASSO
- ▶ LASSO can be used when $p \gg n$

LASSO and Ridge Regression

- ▶ How do we choose λ ? **Cross Validation**
- ▶ When do we use Ridge vs LASSO? When we want a model with less coefficients (predictors) use LASSO
- ▶ LASSO can be used when $p \gg n$
- ▶ Both methods reduce the size of the coefficients but Ridge shrinks to 0 continuously; LASSO shrinks discretely

LASSO and Ridge Regression

- ▶ How do we choose λ ? **Cross Validation**
- ▶ When do we use Ridge vs LASSO? When we want a model with less coefficients (predictors) use LASSO
- ▶ LASSO can be used when $p \gg n$
- ▶ Both methods reduce the size of the coefficients but Ridge shrinks to 0 continuously; LASSO shrinks discretely
- ▶ Both biased in the statistical sense but asymptotically unbiased and can be shown that they select the true features with probability $\rightarrow 1$

LASSO and Ridge Regression

- ▶ Ridge and Lasso can be developed using a Bayesian framework

LASSO and Ridge Regression

- ▶ Ridge and Lasso can be developed using a Bayesian framework
- ▶ Equivalent to setting a Gaussian Prior and Laplace prior on each coefficient for Ridge and Lasso, respectively

LASSO and Ridge Regression

- ▶ Ridge and Lasso can be developed using a Bayesian framework
- ▶ Equivalent to setting a Gaussian Prior and Laplace prior on each coefficient for Ridge and Lasso, respectively
- ▶ Both are convex optimization problems where Ridge has a closed form solution and LASSO must be solved iteratively

LASSO and Ridge Regression

- ▶ Ridge and Lasso can be developed using a Bayesian framework
- ▶ Equivalent to setting a Gaussian Prior and Laplace prior on each coefficient for Ridge and Lasso, respectively
- ▶ Both are convex optimization problems where Ridge has a closed form solution and LASSO must be solved iteratively
- ▶ L1 and L2-Norm can be extended to L-P Norm

LASSO and Ridge Regression

- ▶ Ridge and Lasso can be developed using a Bayesian framework
- ▶ Equivalent to setting a Gaussian Prior and Laplace prior on each coefficient for Ridge and Lasso, respectively
- ▶ Both are convex optimization problems where Ridge has a closed form solution and LASSO must be solved iteratively
- ▶ L1 and L2-Norm can be extended to L-P Norm
- ▶ P-norms < 1 yield sparser models

LASSO and Ridge Regression

- ▶ Ridge and Lasso can be developed using a Bayesian framework
- ▶ Equivalent to setting a Gaussian Prior and Laplace prior on each coefficient for Ridge and Lasso, respectively
- ▶ Both are convex optimization problems where Ridge has a closed form solution and LASSO must be solved iteratively
- ▶ L1 and L2-Norm can be extended to L-P Norm
- ▶ P-norms < 1 yield sparser models
- ▶ Many different ways to regularize parameters, extensive literature exists

LASSO and Ridge Regression

- ▶ Ridge and Lasso can be developed using a Bayesian framework
- ▶ Equivalent to setting a Gaussian Prior and Laplace prior on each coefficient for Ridge and Lasso, respectively
- ▶ Both are convex optimization problems where Ridge has a closed form solution and LASSO must be solved iteratively
- ▶ L1 and L2-Norm can be extended to L-P Norm
- ▶ P-norms < 1 yield sparser models
- ▶ Many different ways to regularize parameters, extensive literature exists
- ▶ Useful in banking, tech, insurance, marketing, and other fields

LASSO and Ridge Regression

- ▶ Amazon Kaggle competition: Employee Access Challenge

LASSO and Ridge Regression

- ▶ Amazon Kaggle competition: Employee Access Challenge
- ▶ Training data on 32,768 employee with

LASSO and Ridge Regression

- ▶ Amazon Kaggle competition: Employee Access Challenge
- ▶ Training data on 32,768 employee with
- ▶ Outcome/Label is binary outcome of whether an employee received access (Average Training Response = 0.94)

LASSO and Ridge Regression

- ▶ Amazon Kaggle competition: Employee Access Challenge
- ▶ Training data on 32,768 employee with
- ▶ Outcome/Label is binary outcome of whether an employee received access (Average Training Response = 0.94)
- ▶ 1,687 teams with most submissions performing reasonably well

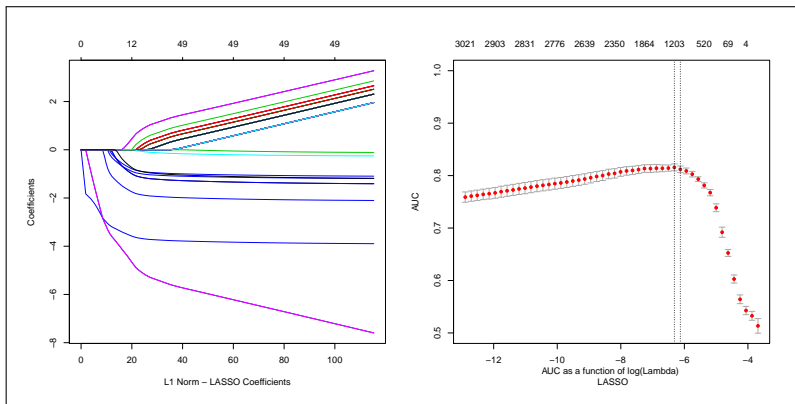
LASSO and Ridge Regression

- ▶ Amazon Kaggle competition: Employee Access Challenge
- ▶ Training data on 32,768 employee with
- ▶ Outcome/Label is binary outcome of whether an employee received access (Average Training Response = 0.94)
- ▶ 1,687 teams with most submissions performing reasonably well
- ▶ Features are all categorical variables and highly sparse

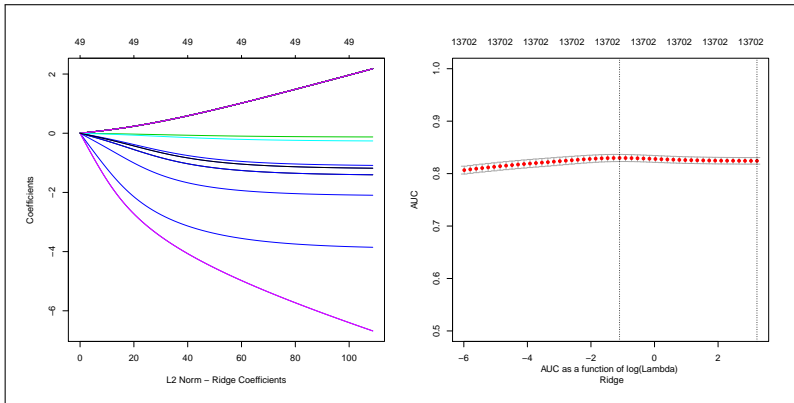
LASSO and Ridge Regression

- ▶ Amazon Kaggle competition: Employee Access Challenge
- ▶ Training data on 32,768 employee with
- ▶ Outcome/Label is binary outcome of whether an employee received access (Average Training Response = 0.94)
- ▶ 1,687 teams with most submissions performing reasonably well
- ▶ Features are all categorical variables and highly sparse
- ▶ Binary representation of all 10 features leads to a very sparse (32,769 × 15,618) matrix

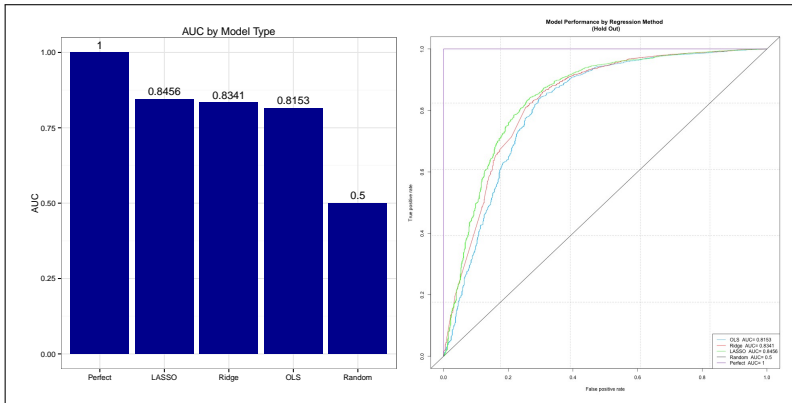
LASSO Model Performance



Ridge Model Performance



Comparison of Ridge, LASSO, and OLS



Useful Links below

Thank you.

- ▶ My Email (Not as Useful)

Useful Links below

Thank you.

- ▶ My Email (Not as Useful)
- ▶ GitHub (Moderately Useful)

Useful Links below

Thank you.

- ▶ My Email (Not as Useful)
- ▶ GitHub (Moderately Useful)
- ▶ GLMNET in R (Very Useful)

Useful Links below

Thank you.

- ▶ My Email (Not as Useful)
- ▶ GitHub (Moderately Useful)
- ▶ GLMNET in R (Very Useful)
- ▶ SKLearn (Also Very Useful)

Useful Links below

Thank you.

- ▶ My Email (Not as Useful)
- ▶ GitHub (Moderately Useful)
- ▶ GLMNET in R (Very Useful)
- ▶ SKLearn (Also Very Useful)
- ▶ Original LASSO Publication (REALLY Useful)

Useful Links below

Thank you.

- ▶ My Email (Not as Useful)
- ▶ GitHub (Moderately Useful)
- ▶ GLMNET in R (Very Useful)
- ▶ SKLearn (Also Very Useful)
- ▶ Original LASSO Publication (REALLY Useful)
- ▶ Original Ridge Publication (Somewhat useful)

Useful Links below

Thank you.

- ▶ My Email (Not as Useful)
- ▶ GitHub (Moderately Useful)
- ▶ GLMNET in R (Very Useful)
- ▶ SKLearn (Also Very Useful)
- ▶ Original LASSO Publication (REALLY Useful)
- ▶ Original Ridge Publication (Somewhat useful)
- ▶ Slides from Stanford (Most Useful)

Useful Links below

Thank you.

- ▶ My Email (Not as Useful)
- ▶ GitHub (Moderately Useful)
- ▶ GLMNET in R (Very Useful)
- ▶ SKLearn (Also Very Useful)
- ▶ Original LASSO Publication (REALLY Useful)
- ▶ Original Ridge Publication (Somewhat useful)
- ▶ Slides from Stanford (Most Useful)
- ▶ Sweet paper on asymptotic properties of Ridge and LASSO