Gandre_Final_Project

Individual Applied Statistics Final Project 40015/50015

Paritosh Gandre

2023-12-10

Group Members : PARITOSH GANDRE

1. (a) Load the package alr4 into memory

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
## Loading required package: car
##
## Loading required package: carData
##
##
## Attaching package: 'car'
##
##
## The following object is masked from 'package:dplyr':
##
##     recode
##
##
## The following object is masked from 'package:purrr':
##
##     some
##
##
## Loading required package: effects
##
## lattice theme set by effectsTheme()
## See ?effectsTheme for details.

## Warning: package 'caret' was built under R version 4.3.2

## Loading required package: lattice
##
## Attaching package: 'caret'
##
```

```
## The following object is masked from 'package:purrr':
##
##     lift
```

(b) From the dataset Downer, construct a dataframe using the variables claved, daysrec, ck, ast, urea, and pcv. Remove any rows with missing data.

(c) Construct a logistic regression model with explanatory variables given in (1b) and outcome as the response. Use an 80/20 split of the data.

(d) Construct a confusion matrix. How accurate is your model?

```
##           predicted
##             0  1
##   died     24  0
##   survived  5  4
```

```
## [1] 0.8484848
```

The accuracy of my model is **0.8484848**.

2. (a) Loading the Dataset:

```
## Warning: package 'glmnet' was built under R version 4.3.2
```

```
## Loading required package: Matrix
```

```
##
## Attaching package: 'Matrix'
```

```
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
```

```
## Loaded glmnet 4.1-8
```

```
## Warning: package 'MASS' was built under R version 4.3.2
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

(b) Lasso Regression Modeling

```
##  [1] 84.27164 77.13000 70.37849 64.05977 58.54201 53.95929 50.16041 47.01171
##  [9] 44.38297 42.02756 39.79265 37.86542 36.25483 34.91941 33.81232 32.89466
## [17] 32.13415 31.51616 31.00690 30.55764 30.16596 29.79392 29.47973 29.21314
## [25] 28.99141 28.77553 28.58393 28.37386 28.12497 27.82560 27.46496 27.14068
## [33] 26.83720 26.57138 26.33243 26.11742 25.92732 25.76846 25.63258 25.49801
## [41] 25.37257 25.18859 24.99289 24.83071 24.69417 24.58642 24.50042 24.42983
## [49] 24.37170 24.32333 24.28339 24.25050 24.22335 24.20098 24.18274 24.16902
## [57] 24.15985 24.15545 24.15363 24.15237 24.15217 24.15329 24.15466 24.15786
## [65] 24.16074 24.16410 24.16720 24.17042 24.17325 24.17623 24.17930 24.18325
## [73] 24.18641 24.18875 24.19057 24.19200
```

(c) Variable Selection:

```
##    crim      zn    chas     nox      rm     dis     rad     tax ptratio   black
##       1       2       4       5       6       8       9      10      11      12
##   lstat
```

```
##      13
```

(d) Evaluation:

(e) Interpretation:

```
## Optimal Lambda: 0.02551743
```

```
## Selected Features: crim zn chas nox rm dis rad tax ptratio black lstat
```

```
## Coefficients of Selected Features: -0.09979684 0.04192955 2.684828 -16.41364 3.859008 -1.40547 0.259
```

```
## Mean Squared Error (CV): 24.15217
```

The fitting of the Lasso regression model to the Boston housing dataset identifies decisive and non-zero important predictors that hint on the fundamental factors responsible for house prices in one of the suburbs of Boston. Positive coefficients will imply features positively related to the house price while a negative relation will be inferred for those coefficients that are negative.The model's predictive capabilities, reflected in performance such as mean squared error metrics, underline its performance while accounting for constraints highlights how nuanced interpretations are needed to identify and follow trends when navigating the intricacies.

3. (a) Load the faithful dataset in R.

(b) Implement polynomial regression models with degrees from 1 to 4.

(c) Use 10-fold cross-validation to compute R2 values for each model.

(d) Identify the degree that corresponds to the highest average crossvalidated R2 value.

Provide the selected degree and its corresponding average R2 value as the solution.

```
## Degree with highest average crossvalidated r2 value is: 4
```

```
## highest average crossvalidated R2 value is: 0.8409718
```