Analysis of Compressive Concrete Strength in Cement

Abhishek Anish

G01460688

1) Project Description

This research project endeavours to construct a predictive model for concrete compressive strength by analysing a comprehensive dataset. The primary focus is on understanding the relationships and patterns among various components involved in concrete mixtures. The central research questions guiding this investigation are twofold: firstly, can an effective model be developed to forecast concrete compressive strength accurately? Secondly, what key factors emerge as significant predictors influencing the compressive strength of concrete? By addressing these questions, the study aims to uncover critical insights into the factors shaping concrete strength, with potential applications in optimizing concrete mixtures for improved structural performance in construction projects.

2) Dataset

	A	B	C	D	E	F	G	H	I
1	Cement (component 1)(kg in a m^3 mixture)	Blast Furnace Slag (component 2)(kg in a m^3 mixture)	Fly Ash (component 3)(kg in a m^3 mixture)	Water (component 4)(kg in a m^3 mixture)	Superplasticizer (component 5)(kg in a m^3 mixture)	Coarse Aggregate (component 6)(kg in a m^3 mixture)	Fine Aggregate (component 7)(kg in a m^3 mixture)	Age (day)	Concrete compressive strength(MPa, megapascals)
2	540.0	0.0	0.0	162.0	2.5	1040.0	676.0	28	79.99
3	540.0	0.0	0.0	162.0	2.5	1055.0	676.0	28	61.89
4	332.5	142.5	0.0	228.0	0.0	932.0	594.0	270	40.27
5	332.5	142.5	0.0	228.0	0.0	932.0	594.0	365	41.05
6	198.6	132.4	0.0	192.0	0.0	978.4	825.5	360	44.30
7	266.0	114.0	0.0	228.0	0.0	932.0	670.0	90	47.03
8	380.0	95.0	0.0	228.0	0.0	932.0	594.0	365	43.70
9	380.0	95.0	0.0	228.0	0.0	932.0	594.0	28	36.45
10	266.0	114.0	0.0	228.0	0.0	932.0	670.0	28	45.85
11	475.0	0.0	0.0	228.0	0.0	932.0	594.0	28	39.29

Figure 1. Original dataset

Dataset Description

The dataset, obtained from the UCI Machine Learning Repository, contains information about the compressive strength of concrete, including details such as the age of the concrete (in days). The quantities of cement, blast furnace slag, fly ash, water, superplasticizer, fine aggregate, and coarse aggregate, expressed in kilograms per cubic meter, are used in the mix. The compressive strength of the concrete is measured in MPa.

Data inspection and preparation of analysis data set

The dataset consists of 1030 observations with 9 columns, encompassing 8 predictor variables and 1 response variable. As part of the pre-processing, column names were altered to enhance usability, streamlining the dataset by excluding component indices and measurement units associated with each component in the original dataset. This adjustment aims to facilitate a more straightforward analysis and interpretation of the data patterns.

3) Research Questions and Exploratory Analysis

The research questions are as follows: -

i. Can a model be built that can predict compressive concrete strength?
ii. What are the key predictors in predicting compressive concrete strength?

Exploratory data analysis was carried out by plotting concrete compressive strength with all the predictors. Correlations between the continuous variables were also analysed.

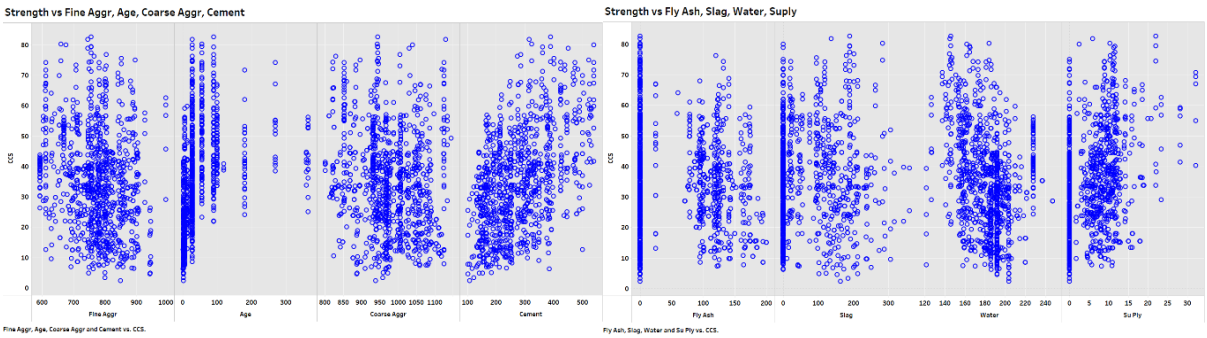

Figure 2: Strength vs Predictors Scatter Plots

Each graph considers 4 predictors simultaneously—one with the first set of 4 predictors and the other with the remaining 4 predictors. Compressive Concrete Strength (CCS) was plotted against all predictors using scatterplots in Tableau software to observe how CCS varies with each predictor. The scatterplot of Strength vs Cement within this analysis reveals a somewhat positive correlation.

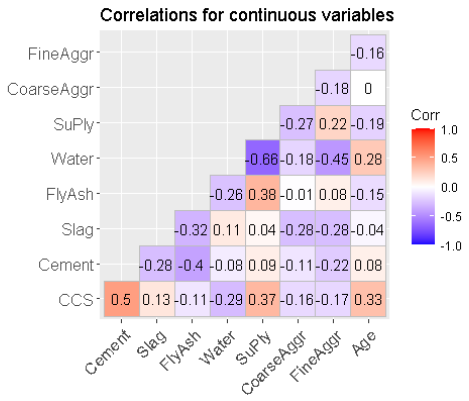

Figure 3. Correlations for continuous variables

Figure 3 shows the correlations between the variables (all of which are continuous). There is a high positive correlation between CCS and cement, while superplasticizer and water also exhibit a high negative correlation. However, since the VIF (variance inflation factor) values were all less than 10 for the linear model using all 8 predictors, none of the predictors were dropped to address multicollinearity.

4) Data analysis

The software used includes Tableau for the initial plots, R programming through the R Studio IDE for the analysis. The packages used are: MASS, dplyr, rpart, rpart.plot, DMwR2, tidyverse, glmnet, ggcorrplot, GGally, leaps, gridExtra, coefplot, car.

The research questions were addressed through a comprehensive analysis employing various regression techniques, including linear regression, lasso regression, and regression tree, utilizing the R programming language within the R Studio environment. For linear regression, both best subset selection and AIC-based methods were applied to optimize the model. To enhance interpretability, the concrete compressive strength (CCS) variable was renamed as the response variable.

I. A simple linear model was carried out using all the 8 predictors.

```
Call:
lm(formula = response ~ ., data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-28.654  -6.302   0.703   6.569  34.450

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -23.331214  26.585504  -0.878 0.380372
Cement        0.119804   0.008489  14.113  < 2e-16 ***
Slag          0.103866   0.010136  10.247  < 2e-16 ***
FlyAsh        0.087934   0.012583   6.988 5.02e-12 ***
Water        -0.149918   0.040177  -3.731 0.000201 ***
SuPly         0.292225   0.093424   3.128 0.001810 **
CoarseAggr    0.018086   0.009392   1.926 0.054425 .
FineAggr      0.020190   0.010702   1.887 0.059491 .
Age           0.114222   0.005427  21.046  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.4 on 1021 degrees of freedom
Multiple R-squared:  0.6155,    Adjusted R-squared:  0.6125
F-statistic: 204.3 on 8 and 1021 DF,  p-value: < 2.2e-16
```

Figure 4. Summary of linear model using all the predictors

Through figure 4, the significance of Coarse Aggregate and Fine Aggregate can be noted as borderline compared to the remaining predictors. However, the overall p-value for the model is extremely low (2.2e-16), indicating that the model as a whole is highly significant. This suggests that, collectively, the predictors have a significant impact on the variation in Compressive Concrete Strength. The R-squared value of 0.6155 indicates that approximately 61.55% of the variability in the response variable is explained by the model.

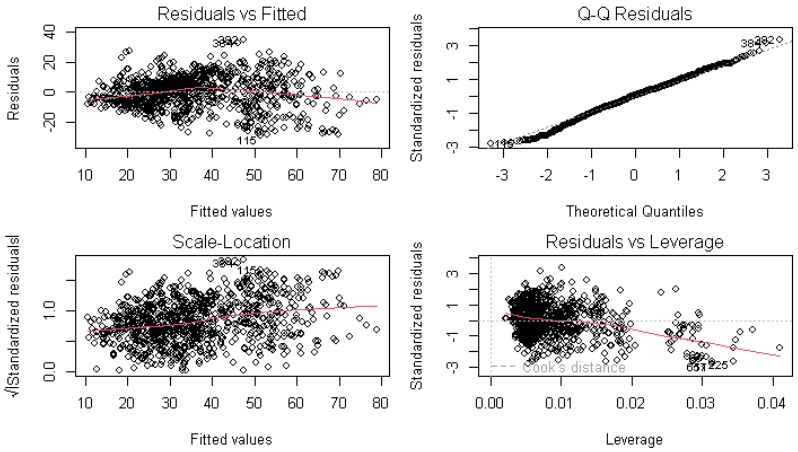

Figure 5. Residual vs Fitted, Q-Q Residuals, Scale- Location and residuals vs Leverage plots for the model using all the predictors.

Figure 5 suggests a reasonable degree of linearity between the response and predictor, as evidenced by the residuals vs. fitted plots, which show an approximately horizontal red line. It's worth noting that while the red line is not perfectly horizontal, the deviation isn't much. However, the Q-Q residuals plot indicates a departure from normality, with several values deviating from the ideal normal line. On a positive note, there is evidence of equal error variances, depicted by a horizontal red line in the scale-location plot. Additionally, the residuals vs. fitted values plot exhibits a pipe or football-shaped

pattern (spread out) rather than a fan-shaped figure (clustered at one end and spread out at the other end), suggesting that the variance of the residuals is relatively constant across the range of fitted values.

II. Linear regression models using best subset methods

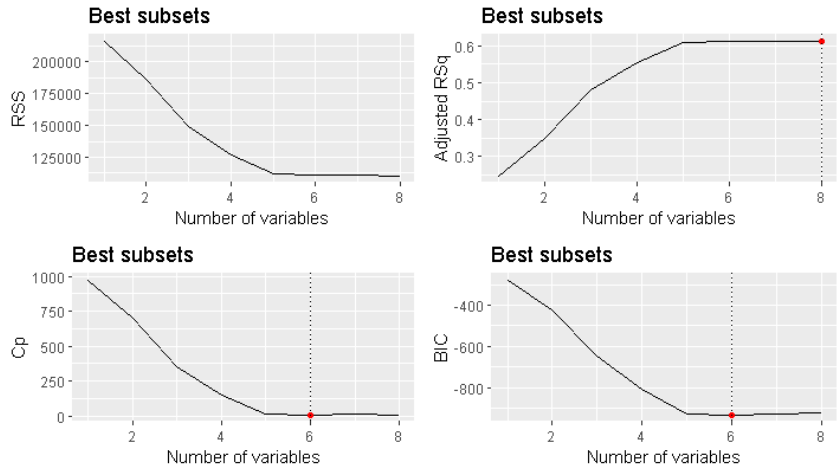

Figure 6. Plot between the number of variables and various statistical metrics (Residual Sum of Squares, Adjusted R-Squared, Mallows' Cp, and Bayesian Information Criterion).

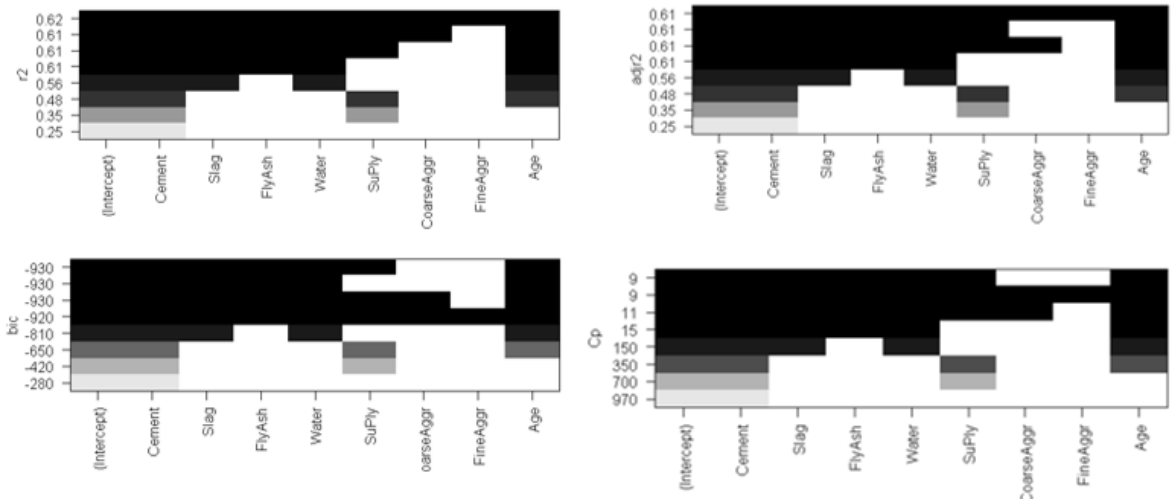

Figure 7 shows the various accuracy measures vs selected predictors

Figures 6 and 7 illustrate that the number of predictors can either be 6 or 8. Moreover, it is evident that coarse aggregate and fine aggregate do not play a significant role when 6 predictors are utilized. Additionally, it should be noted that the values of various statistical metrics do not vary significantly.

```
> round(coef(regfit.full,6), 3)
(Intercept)      Cement        Slag      FlyAsh       Water       SuPly         Age
     28.993       0.105       0.086       0.069      -0.218       0.240       0.113
> round(coef(regfit.full,8), 3)
(Intercept)      Cement        Slag      FlyAsh       Water       SuPly  CoarseAggr
    -23.331       0.120       0.104       0.088      -0.150       0.292       0.018
   FineAggr         Age
      0.020       0.114
```

Figure 8. Shows the predictors and coefficients for best 6 and 8 variable subsets

Using a validation set to choose between model indicated 5 predictors (excluding superplasticizer, coarse aggregate and fine aggregate) which gave the smallest test set MSE.

```
> round(coef(reg.best,5),3)
(Intercept)      Cement        Slag      FlyAsh       Water         Age
     34.826       0.110       0.092       0.080      -0.255       0.114
>
```

Figure 9. Shows the predictors and coefficients for best 5 variable subsets

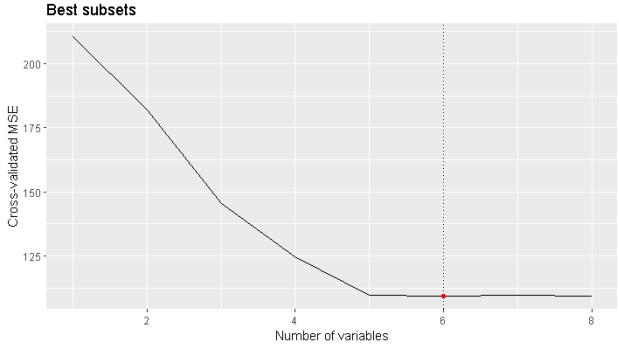

Figure 10. Number of variables vs Cross Validated MSE (K cross validation)

K cross validation shows (using k as 10) 6 predictors as the most relevant which gave the smallest Cross validated MSE.

The above analysis involved the generation of subsets with varying numbers of predictors, ranging from 1 to 8, through an exhaustive search algorithm. Performance measures such as R-squared, residual sum of squares (RSS), adjusted R-squared, Cp, and BIC were computed for each subset. Visualizations, including line plots and scatter plots, were created to illustrate the results, showcasing the trade-offs and characteristics of the different subsets. The coefficients for the best 6- and 8-variable subsets were displayed, revealing the impact of variable selection on the model. Training-set mean squared errors (MSE) were calculated for these models. Subsequently, a validation set was employed to choose among the models, leading to the identification of the best 5-variable model based on test MSE. Cross-validation was then employed to further assess model performance, with 10-fold cross-validation revealing the best 6-variable model in terms of minimized cross-validated MSE. The coefficients for this model were also computed, providing insights into the selected predictors' contributions.

III. Lasso regression

Lasso regression was carried out in order to identify the optimal regularization parameter (lambda) that minimizes mean squared error through cross-validation on a training set, facilitating variable selection and assessing the predictive performance of the model on both the training and test datasets.

```
> round(lasso.coef.min[lasso.coef.min != 0], 3)
(Intercept)      Cement        Slag      FlyAsh       Water       SuPly  CoarseAggr
    -13.177       0.117       0.100       0.083      -0.162       0.287       0.015
   FineAggr         Age
      0.016       0.114
>
> # Number of predictors with non-zero coefficients for best lambda
> (nvars <- length(lasso.coef.min[lasso.coef.min!=0])-1)
[1] 8
```

Figure 11. Number of predictors and their non-zero coefficients for best lambda – Lasso regression

```
> lasso.coef.1se <- predict(out,type="coefficients",s=bestlam.1se)[1:(ncol(x)+1),]
> round(lasso.coef.1se[lasso.coef.1se!=0], 3)
(Intercept)      Cement        Slag      FlyAsh       Water       SuPly    FineAggr
     32.309       0.086       0.063       0.033      -0.179       0.411      -0.002
         Age
      0.100
>
> # Number of predictors with non-zero coefficients for 1se rule
> (nvars <- length(lasso.coef.1se[lasso.coef.1se!=0])-1)
[1] 7
```

Figure 12. Number of predictors and their non-zero coefficients for lambda 1se – Lasso regression

In the Lasso regression analysis, the dataset was divided into training and test sets, and the mean sum-of-squares for the test set was recalculated for subsequent R-squared computations. Lasso regression was performed over a range of 100 lambda values, and cross-validation on the training set identified the optimal lambda values as 0.008 and 0.528, respectively, for minimizing mean squared error (MSE) and following the 1-standard-error rule. Plots were generated to visualize the lambda values and corresponding coefficient paths. Evaluation on the test set for the optimal lambda values revealed a test MSE of 113.5209, RMSE of 10.65462, and an R-squared of 0.5834. Residuals versus fitted values were illustrated using a scatter plot. A similar analysis for the 1-standard-error rule resulted in a test MSE of 114.7777, RMSE of 10.71343, and an R-squared of 0.5788. Finally, applying the optimal lambda values to the full dataset yielded non-zero coefficients for eight predictors, including 'Cement,' 'Slag,' 'FlyAsh,' 'Water,' 'SuPly,' 'CoarseAggr,' 'FineAggr,' and 'Age.' The coefficient plot for this model was generated using the coefplot package. For the 1-standard-error rule, seven predictors showed non-zero coefficients, excluding 'CoarseAggr.' The respective coefficient plot was also created. These findings contribute valuable insights into the variable selection process and the impact of regularization in the Lasso regression context.

```
# Results so far:
#                        # predictors  training-set MSE   Test MSE
#   Best BIC & Cp model:       6          107.6148
#   Best Adj Rsq model:        8          107.1972
#
#   Validation set:
#   Best test-set model        5                          111.8174
#   K-fold CV best model       6                          109.1922
#
#   Best-lambda Lasso          8                          113.5209
#   1se-lambda Lasso           7                          114.7777
```

Figure 13. Results table including number of predictors, training set MSEs and test MSEs

IV. Linear Regression using Stepwise search for minimum AIC models

Linear Regression using Stepwise search for minimum AIC models was employed to include interactions and identify the most influential predictors in the dataset. The model was validated using a test set, and performance metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared were calculated, demonstrating the model's effectiveness in predicting the response variable.

```
> (coef <- round(coef(model.final), 5))
      (Intercept)            Cement              Slag            FlyAsh             Water
        -307.73536           0.47072          -0.25695          -0.03691           1.65414
             SuPly        CoarseAggr         FineAggr               Age       Cement:Slag
           1.96411           0.26005          -0.04170          -0.56812           0.00020
    Cement:FlyAsh      Cement:Water     Cement:SuPly Cement:CoarseAggr       Cement:Age
           0.00029          -0.00157          -0.00496          -0.00009           0.00057
     Slag:FlyAsh   Slag:CoarseAggr     Slag:FineAggr          Slag:Age      FlyAsh:Water
           0.00044           0.00003           0.00030           0.00080          -0.00153
    FlyAsh:SuPly   FlyAsh:FineAggr       FlyAsh:Age  Water:CoarseAggr        SuPly:Age
          -0.00680           0.00034           0.00180          -0.00125           0.00659
    FineAggr:Age
           0.00057
>
>
> # number of effects (other than the intercept) in the final model
> (length(coef) -1)
[1] 25
```

Figure 14. Final model coefficients and interactions using stepwise search for minimum AIC model

A stepwise search for the minimum AIC model was conducted using the entire dataset, leading to a model with 30 coefficients. The resulting coefficients provide insights into the selected predictors and their corresponding weights. A comparison of AIC values between the initial and stepwise models indicates a lower AIC for the stepwise model, suggesting an enhanced model fit. Subsequently, a validation-set approach was employed to estimate the Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and test-set R-squared. Notably, the stepwise model exhibited the lowest MSE and RMSE values, indicating superior predictive performance compared to all previously analysed linear models. The test-set R-squared for this model was also the highest, signifying robust explanatory power. The final step involved building the ultimate model using all available data, resulting in a model with 25 effects (comprising 8 predictors and 17 interactions, excluding the intercept). The coefficients for this comprehensive model are presented, encapsulating the refined understanding of the relationships within the dataset.

V. Regression Tree

Regression tree was used to model the relationship between the response variable, concrete compressive strength, and a set of predictor variables, including Age, Cement, CoarseAggr, FineAggr, Slag, SuPly, and Water (excluding Flyash) utilizing the rpart package.

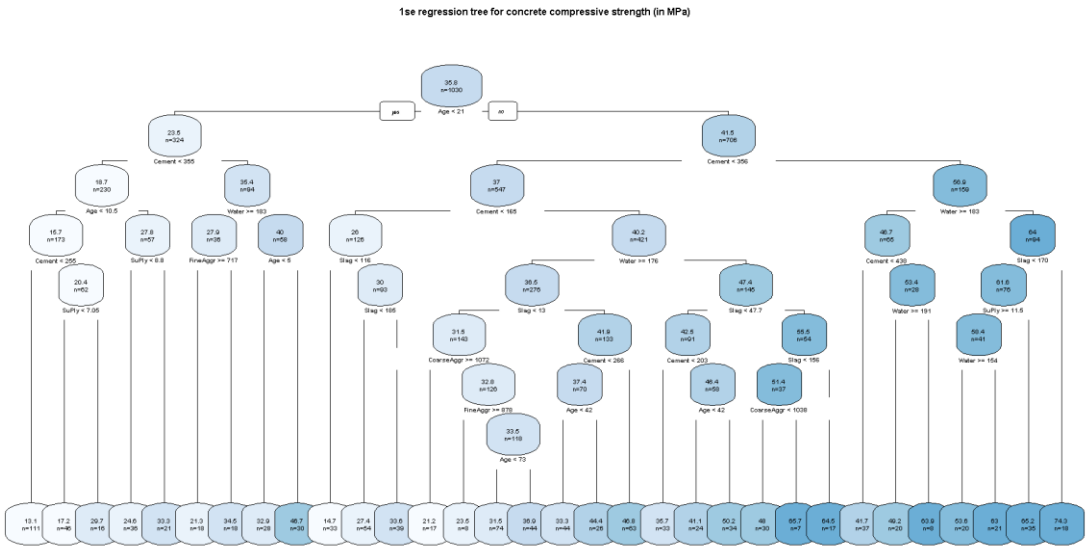

Figure 15. Regression tree built by the algorithm.

```
Variables actually used in tree construction:
[1] Age        Cement     CoarseAggr FineAggr   Slag       SuPly      Water

Root node error: 287175/1030 = 278.81

n= 1030
```

Figure 16. Model summary output indicating predictors used for regression tree

```
#  Results:                                MSE          RMSE
#     Cross-validated MSE                67.00641     8.185744
#     Validation set MSE (set.seed(7))   68.21776     8.259404
#     Validation set MSE (set.seed(5))   64.34217     8.021357
```

Figure 17. Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) for Cross-Validation and Different Validation Sets

Figure 15,16 and 17 shows the 1se (1 standard error) pruned regression tree, model summary to indicate the predictors used and the MSE and RMSE for cross validation and different validation sets respectively.

The initial tree was pruned based on the complexity parameter (cp), and a 1-standard-error rule was applied to obtain a pruned regression tree model. The variables involved in tree construction include Age, Cement, CoarseAggr, FineAggr, Slag, SuPly, and Water. The root node error was found to be 278.81, indicating the initial misclassification rate. The cross-validated Mean Squared Error (MSE) for the pruned tree was computed as 67.00641, with a corresponding Root Mean Squared Error (RMSE) of 8.185744. Additionally, validation set experiments with different seed values resulted in MSE values of 68.21776 and 64.34217, with corresponding RMSE values of 8.259404 and 8.021357, respectively. These results provide insights into the predictive performance of the regression tree model, indicating its ability to generalize to new data and predict concrete compressive strength effectively.

5) Results and Conclusion

After applying various regression models to predict concrete compressive strength (CCS), the results revealed that the best linear regression model, obtained through a stepwise search for the Minimum AIC model, involved all 8 predictors and 25 effects, including interactions. This model exhibited a Mean Squared Error (MSE) of 83.71, a Root Mean Squared Error (RMSE) of 9.14, and an R-squared value of 0.69. The Lasso regression model, utilizing all 8 predictors, demonstrated a slightly higher MSE of 113.52, RMSE of 10.65, and R-squared of 0.58. Additionally, the regression tree model, cross-validated during training, showcased an MSE of 67.006 and RMSE of 8.19. Remarkably, the regression tree model, which used 7 predictors (excluding Fly Ash), revealed valuable insights into the influential factors for obtaining the highest CCS, indicating an average of 74.3. Notably, these influential factors included an age greater than 21 days, cement content exceeding 356 kilograms per cubic meter, water content less than or equal to 183 kilograms per cubic meter, and blast furnace slag content surpassing 170 kilograms per cubic meter. This comprehensive analysis provides valuable information for understanding and predicting concrete compressive strength, incorporating the strengths of different regression methodologies.

6) Future work

Future research includes refining and extending the current models, considering a broader spectrum of factors, and leveraging cutting-edge methodologies to address the evolving challenges in the field of concrete strength prediction. This encompasses exploring techniques to handle non-linearity, such as incorporating polynomial regression models, to capture more complex relationships between predictors and concrete compressive strength.

References

1. Dataset: UCI Machine Learning Repository. (n.d.). Concrete Compressive Strength. Retrieved November 15, 2023, from https://archive.ics.uci.edu/dataset/165/concrete+compressive+strength
2. RStudio: RStudio Team. (2023). RStudio Desktop IDE (Version 2023.06.0-421) [Computer software]. PBC. Retrieved from https://posit.co/download/rstudio-desktop/
3. Tableau: Tableau. (n.d.). Retrieved December 10, 2023, from https://www.tableau.com/
4. MASS: Venables, W. N. & Ripley, B. D. (2002) Modern Applied Statistics with S. FourthEdition. Springer, New York. ISBN 0-387-95457-0

5. car: Fox J, Weisberg S (2019). _An R Companion to Applied Regression_, Third edition. Sage, Thousand Oaks CA. https://socialsciences.mcmaster.ca/jfox/Books/Companion/
6. dplyr: Wickham H, François R, Henry L, Müller K, Vaughan D (2023). _dplyr: A Grammar of Data Manipulation_. R package version 1.1.2, https://CRAN.R-project.org/package=dplyr
7. rpart: Therneau T, Atkinson B (2022). _rpart: Recursive Partitioning and Regression Trees_. R package version 4.1.19, https://CRAN.R-project.org/package=rpart
8. rpart.plot: Milborrow S (2022). _rpart.plot: Plot 'rpart' Models: An Enhanced Version of
9. 'plot.rpart'_. R package version 3.1.1, https://CRAN.R-project.org/package=rpart.plot
10. DMwR2: Torgo, L. (2016). Data Mining with R, learning with case studies, 2nd editionChapman and Hall/CRC. URL: http://ltorgo.github.io/DMwR2
11. tidyverse: Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H (2019).“Welcome to the tidyverse.” _Journal of Open Source Software_, *4*(43), 1686. doi:10.21105/joss.01686, https://doi.org/10.21105/joss.01686
12. glmnet: Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software, 33(1), 1–22. https://doi.org/10.18637/jss.v033.i01
13. ggcorrplot: Kassambara, A., & Mundt, F. (2023). _ggcorrplot: Visualization of a Correlation Matrix using 'ggplot2'. R package version 0.1.4.1 https://CRAN.R-project.org/package=ggcorrplot
14. GGally: Schloerke, B., Crowley, J., Cook, D., Briatte, F., Marbach, M., Thoen, E., & Elberg, A. (2021). GGally: Extension to 'ggplot2'. R package version 2.1.2. https://CRAN.R-project.org/package=GGally
15. leaps: Miller TLboFcbA (2020). leaps: Regression Subset Selection. R package version 3.1. https://CRAN.R-project.org/package=leaps
16. gridExtra: Auguie, B. (2017). gridExtra: Miscellaneous Functions for "Grid" Graphics. R package version 2.3. https://CRAN.R-project.org/package=gridExtra
17. coefplot: Lander, J. P. (2021). coefplot: Plots Coefficients from Fitted Models. R package version 1.2.8. https://CRAN.R-project.org/package=coefplot
18. Sigman, R. (2023). subset_selection_hitters_data.R [R code]. Downloaded from Blackboard site for STAT 515, Section 004, Fall semester, 2023, on November 25 2023.
19. Sigman, R. (2023). stepaic_example.R [R code]. Downloaded from Blackboard site for STAT 515, Section 004, Fall semester, 2023, on November 25 2023.
20. Sigman, R. (2023). 3_Decision_trees_v2.R [R code]. Downloaded from Blackboard site for STAT 515, Section 004, Fall semester, 2023, on November 25 2023.
21. Sigman, R. (2023). rss_regress_funcs_v5.R [R code]. Downloaded from Blackboard site for STAT 515, Section 004, Fall semester, 2023, on October 24 2023.

APPENDIX

The appendix includes the column name changes done for easy usability and the collective results of our research group.

Original column	Altered column name
Cement (component 1)(kg in a m^3 mixture)	Cement
Blast Furnace Slag (component 2) (kg in a m^3 mixture)	Slag
Fly Ash (component 3)(kg in a m^3 mixture)	FlyAsh
Water (component 4)(kg in a m^3 mixture)	Water
Superplasticizer (component 5)(kg in a m^3 mixture)	SuPly
Coarse Aggregate (component 6)(kg in a m^3 mixture)	CoarseAggr
Fine Aggregate (component 7)(kg in a m^3 mixture)	FineAggr
Age (day)	Age
Concrete compressive strength(MPa, megapascals)	CCS

Table 1. Column name alterations

Abhishek Anish		Gautham Krishna	
Model	Result	Model	Result
Multilinear Regression	MSE – 83.71 RMSE – 9.14 R squared – 0.69	Logistic Regression	pseudo-R-squared – 0.5642 AIC – 94.493 p value (6.75e -18) Accuracy – 0.9689
Lasso Regression	MSE – 113.52 RMSE – 10.65 R squared – 0.58	Classification Tree	cross validated training – 0.203 Root node error – 0.559 Classification Error - 0.2582
Regression Tree	MSE – 67.006 RMSE – 8.19	Random Forest	mtry = 8 R squared – 0.8825 MSE – 27.57 RMSE – 5.25

Table 2. Collective results