**Final project**

Hung Cao  
Cody Hudepohl  
Trang Le  
Duy Tran

**Table of Contents**

I. Background 2

II. Data Format 2

III. Goal and Methods of Analyzing Data 2

1. Goal 2

2. Methods 3

IV. Explore Data Analysis………………………………………………………………………………………………………………..3

1. Summary Table 3

2. Statistical Analysis Graphs 4

V. Variable Selection and Results for Calwpct………………………………………………………………………….7

1. Linear Regression 7

2. Forward Selection 8

3. Backward Selection 8

4. Stepwise Selection 9

5. LASSO 9

a. LASSO Min .9

b. LASSO 1SE 10

6. Regression Tree and Pruned Tree 10

a. Regression Tree 10

b. Pruned Tree 11

7. Results and Comparison 13

a. Results 13

b. Comparison 13

8. Conclusion for CalWorks assistance program 14

VI. Variable Selection and Results for Mealpct……………………………………………………………………….14

1. Linear Regression 14

2. Forward Selection 15

3. Backward Selection 15

4. Stepwise Selection 16

5. LASSO 16

a. LASSO Min 16

b. LASSO 1SE 17

6. Regression Tree and Pruned Tree 17

a. Regression Tree 17

b. Pruned Tree 18

7. Results and Comparison 20

a. Results 20

b. Comparison 21

8. Conclusion for Price-Reduced Lunch program 21

**I. Background**

This dataset is conducted by California Department of Education. It gathers 17 descriptive variables from 420 schools. The data’s main usage is to determine the level of public fund that a school may receive. The data we have is only from 1998 to 1999.

There are several public funds that student might receive, and they are recorded in this dataset: CalWorks, reduced-price lunch, and expenditure. According to Wikipedia, “CalWork is a welfare program that gives cash aid and services to eligible needy California families.”

**II. Data Format**

The information from the schools included in this dataset are:

* ~~Distcod: district code~~
* ~~County: name of county~~
* ~~District: name of district~~
* Grspan: grade span of district
* Enrltot: total enrollment
* Teachers: number of teachers
* **Calwpct: percent qualifying for CalWorks** (responsive variable)
* **Mealpct: percent qualifying for reduced-price lunch** (responsive variable)
* Computer: number of computers
* Testscr: average test score (reading + math)/2
* Compstu: number of computers per student
* Expnstu: expenditure per student
* Str: student-to-teacher ratio
* Avginc: district average income
* Elpct: percent of English learners
* Readscr: average reading score
* Mathscr: average math score

- CalWorks: CalWORKs is a public assistance program that provides cash aid and services to eligible families that have a child(ren) in the home. The program serves all 58 counties in the state and is operated locally by county welfare departments.

- Reduced Price Lunch: a public assistance program that provides subsidised or free lunch for under-privileged students

* We take out the first 3 explanatory variables since district/county codes and names really does not impact the percent qualifying for CalWorks and Reduced Price Lunch.

**III. Goal and Methods of Analyzing Data**

1. Goal

- Our goal for this project is to determine the level of public funding that a school may receive based off all variables in dataset such as school average test score, number of teachers, district average income and so on.

- The reason why we choose this goal is because we would like to understand what really affects funding levels of the public programs. We would like to see whether test score or average district income have a better influence on the percent qualifying for those public assistance programs.

2. Methods

- Since our responsive variables is continuous, we will use different variable selection methods to analyze this data such as:

* Classical Method
  + Linear Regression
  + Forward/Backward/Stepwise using AIC
* Regularized Method
  + Least Absolute Shrinkage and Selection Operator (LASSO)
    - LASSO MIN
    - LASSO 1SE
* Regression Tree and Pruned Tree

**IV. Explore Data Analysis**

1. Summary Tables

Taking an overview look at the dataset, we run the command head(Caschool).

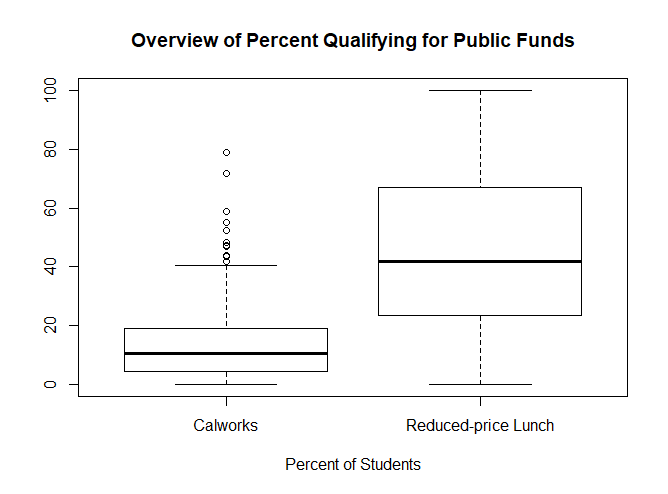
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **grspan** | **enrltot** | **teachers** | **calwpct** | **mealpct** | **computer** | **testscr** |
| KK-08 | 195 | 10.9 | 0.5102 | 2.0408 | 67 | 690.8 |
| KK-08 | 240 | 11.15 | 15.4167 | 47.9167 | 101 | 661.2 |
| KK-08 | 1550 | 82.9 | 55.0323 | 76.3226 | 169 | 643.6 |
| KK-08 | 243 | 14 | 36.4754 | 77.0492 | 85 | 647.7 |
| KK-08 | 1335 | 71.5 | 33.1086 | 78.427 | 171 | 640.85 |
| KK-08 | 137 | 6.4 | 12.3188 | 86.9565 | 25 | 605.55 |
| **compstu** | **expnstu** | **str** | **avginc** | **elpct** | **readscr** | **mathscr** |
| 0.34359 | 6384.911 | 17.88991 | 22.69 | 0 | 691.6 | 690 |
| 0.420833 | 5099.381 | 21.52466 | 9.824 | 4.583333 | 660.5 | 661.9 |
| 0.109032 | 5501.955 | 18.69723 | 8.978 | 30 | 636.3 | 650.9 |
| 0.349794 | 7101.831 | 17.35714 | 8.978 | 0 | 651.9 | 643.5 |
| 0.12809 | 5235.988 | 18.67133 | 9.080333 | 13.85768 | 641.8 | 639.9 |
| 0.182482 | 5580.147 | 21.40625 | 10.415 | 12.40876 | 605.7 | 605.4 |

Then we find the summary of each variable in the dataset, which includes its minimum value, quartiles, median, mean, and maximum value of some of the variables that we are interested in

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **CalPct** | **MealPct** | **Avg Inc** | **Expnstu** | **TestScr** |
| Min | 0 | 0 | 5.335 | 3926 | 605.5 |
| 1st Qt | 4.395 | 23.28 | 10.639 | 4906 | 640 |
| Meadian | 10.52 | 41.75 | 13.728 | 5215 | 654.5 |
| Mean | 13.246 | 44.71 | 15.317 | 5312 | 654.2 |
| 3rd Qt | 18.981 | 66.86 | 17.629 | 5601 | 666.7 |
| Max | 78.994 | 100 | 55.328 | 7712 | 706.8 |

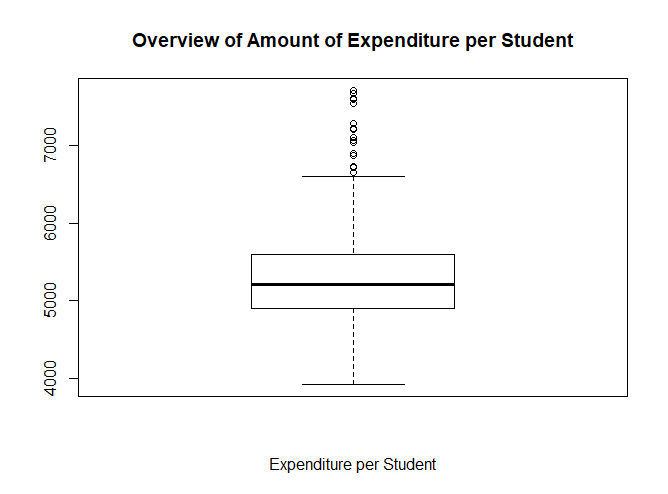
2. Statistical Analysis Graphs

Since we want to determine the level of public fund that a school may receive, we drew a boxplot to take a look at the overview of the percentage of students which qualify for each public fund:



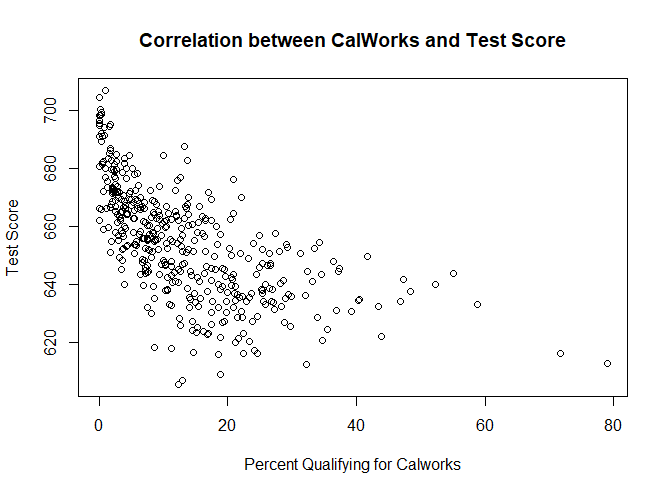
As the boxplot shows, majority of school districts qualify only 5% to 40% of their students for this program with the average of 10%. There are only a few outlier schools which qualify over 40% of their student for CalWorks. Furthermore, the amount of students which qualify for this program is much higher than CalWorks with the range of 5% to 100%. The average is 40$, which is much higher than 10% of CalWorks.

We then draw another boxplot to take a look at the amount of expenditure that each student may receive.



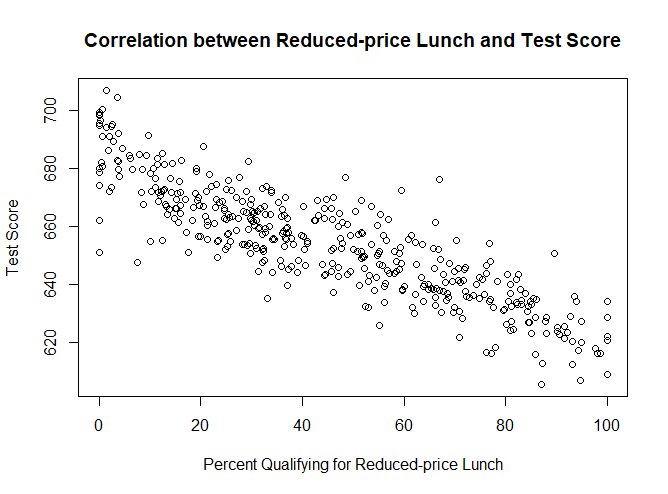
From the boxplot, most students receive from $4,000 to $6,700. There are some outliers of students receiving more than $6,700 up to $7,500.

We want to know how these public funds correlate with how well the students perform academically. Therefore, we drew some scatterplots to demonstrate this relation.



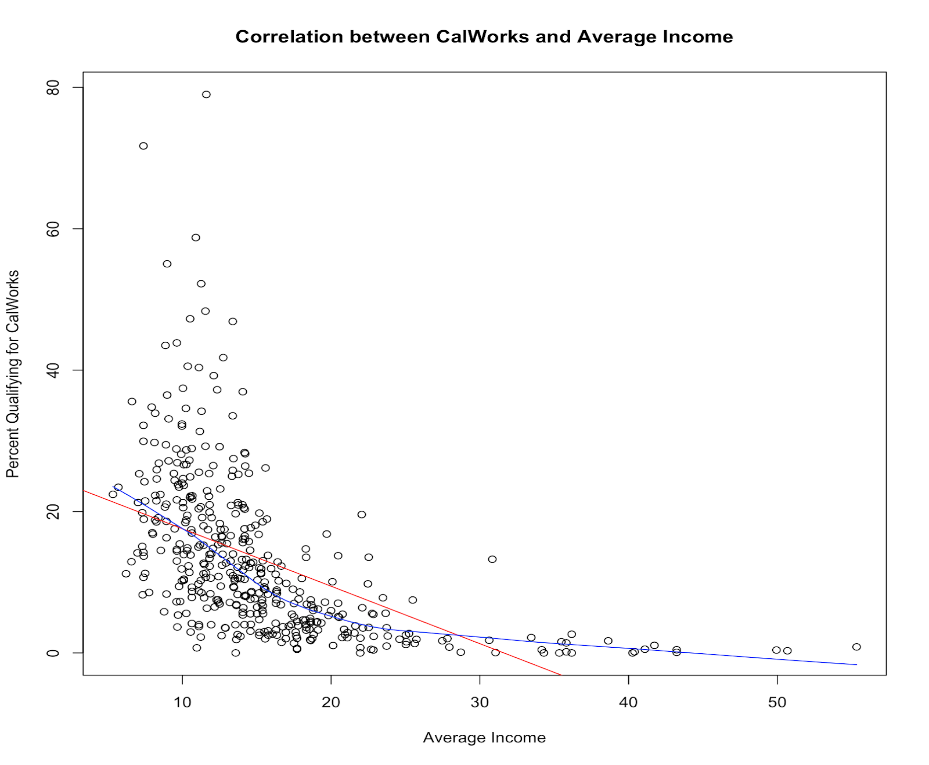
The graph skews to the left, which means students majority of students having average test score qualifies for CalWorks. It is quite surprising that students with high test score have the lowest qualifying percentage for Calworks. We will further analyze the data in our dataset to really understand and come up with a reason for this surprise finding.

The next graph we want to take a look at is the correlation between percentage of students qualifying for Reduce-price Lunch and their test score.

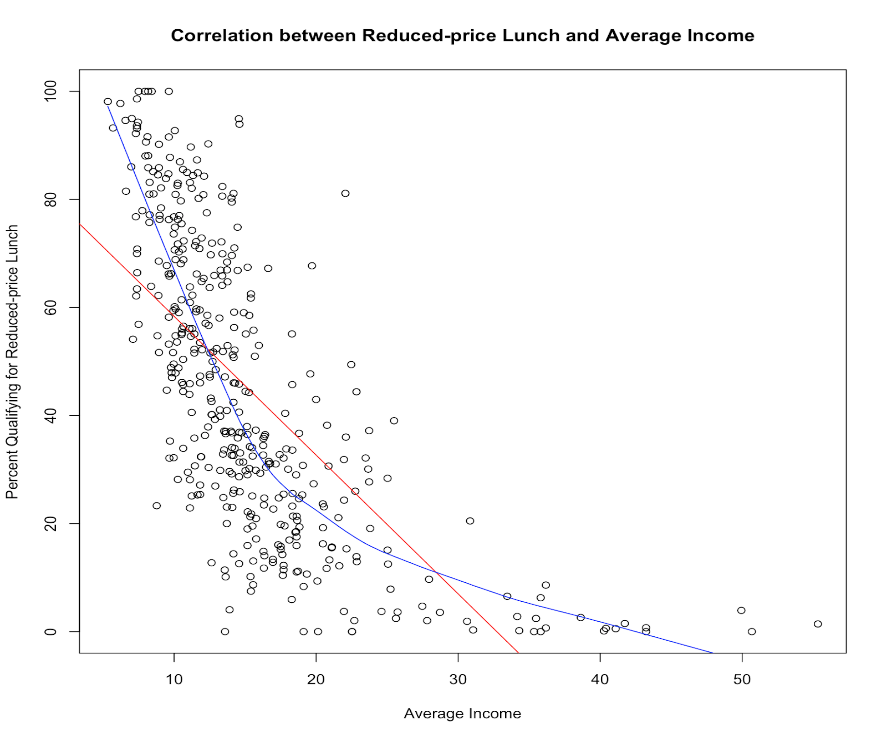


Again, from the graph, we see a negative collreation strongly appear between reduced price lunch and student test score.

Lastly, since we also interested in the correlation between average Income Vs percent qualified in both programs. Here are graphs to indicate those relationship.



Looking at the above graph, the lower your average level of income the higher you can qualifying for Calworks programs. This is also true in the case of Reduced-Price lunch which mean the lower your average income level the higher chance you can be qualified for the program



**V. Variable Selection and Results for CalWorks**

For this part, we run 8 different models in R and also calculate all the associates performance measurement figures for each model. Below is the summary result of model selections for each model. Those highlighted variables are those that is significant in each model.

**1. Linear Regression**

Here is the result we get after running the linear regression.

Call:

lm(formula = calwpct ~ ., data = data.train)

Residuals:

Min 1Q Median 3Q Max

-17.299 -3.316 -0.799 2.366 44.458

Coefficients: (1 not defined because of singularities)

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.032e+01 3.057e+01 0.338 0.735789

grspanKK-08 4.765e-02 1.011e+00 0.047 0.962422

enrltot -2.779e-03 1.474e-03 -1.885 0.060343 .

teachers 5.326e-02 3.185e-02 1.673 0.095373 .

**mealpct 3.561e-01 3.039e-02 11.719 < 2e-16 \*\*\***

computer 4.163e-03 2.853e-03 1.459 0.145474

testscr -1.792e-01 1.033e-01 -1.735 0.083632 .

compstu -1.033e+01 6.515e+00 -1.586 0.113775

**expnstu 3.016e-03 8.245e-04 3.658 0.000296 \*\*\***

str 2.603e-01 2.888e-01 0.901 0.368149

avginc -6.000e-02 8.813e-02 -0.681 0.496439

**elpct -2.009e-01 3.118e-02 -6.445 4.2e-10 \*\*\***

readscr 1.335e-01 1.084e-01 1.232 0.218784

mathscr NA NA NA NA

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.49 on 323 degrees of freedom

Multiple R-squared: 0.6623, Adjusted R-squared: 0.6497

F-statistic: 52.78 on 12 and 323 DF, p-value: < 2.2e-16

**2. Forward Selection**

Call:

lm(formula = calwpct ~ mealpct + elpct + expnstu + computer +

mathscr + compstu, data = data.train)

Residuals:

Min 1Q Median 3Q Max

-17.870 -3.137 -0.692 2.413 44.997

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 31.7416454 22.2130789 1.429 0.153964

**mealpct 0.3561204 0.0244158 14.586 < 2e-16 \*\*\***

**elpct -0.2184277 0.0269709 -8.099 1.09e-14 \*\*\***

**expnstu 0.0027645 0.0005993 4.613 5.69e-06 \*\*\***

**computer 0.0029127 0.0008671 3.359 0.000874 \*\*\***

mathscr -0.0696774 0.0333254 -2.091 0.037311 \*

compstu -9.0534787 5.6707960 -1.597 0.111335

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.479 on 329 degrees of freedom

Multiple R-squared: 0.6572, Adjusted R-squared: 0.651

F-statistic: 105.1 on 6 and 329 DF, p-value: < 2.2e-16

**3. Backward Selection**

Call:

lm(formula = calwpct ~ enrltot + teachers + mealpct + computer +

testscr + compstu + expnstu + elpct, data = data.train)

Residuals:

Min 1Q Median 3Q Max

-17.888 -3.246 -0.682 2.319 44.995

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 3.425e+01 2.619e+01 1.308 0.1919

enrltot -2.145e-03 1.290e-03 -1.662 0.0974 .

teachers 4.012e-02 2.807e-02 1.429 0.1539

**mealpct** 3.543e-01 2.699e-02 13.125 < 2e-16 \*\*\*

computer 4.202e-03 2.795e-03 1.504 0.1337

testscr -7.188e-02 3.948e-02 -1.820 0.0696 .

compstu -1.069e+01 6.382e+00 -1.676 0.0948 .

**expnstu** 2.640e-03 6.266e-04 4.213 3.26e-05 \*\*\*

**elpct** -2.187e-01 2.767e-02 -7.904 4.18e-14 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.478 on 327 degrees of freedom

Multiple R-squared: 0.6594, Adjusted R-squared: 0.6511

F-statistic: 79.13 on 8 and 327 DF, p-value: < 2.2e-16

**4. Stepwise Selection**

Call:

lm(formula = calwpct ~ mealpct + elpct + expnstu + computer +

mathscr + compstu, data = data.train)

Residuals:

Min 1Q Median 3Q Max

-17.870 -3.137 -0.692 2.413 44.997

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 31.7416454 22.2130789 1.429 0.153964

**mealpct 0.3561204 0.0244158 14.586 < 2e-16 \*\*\***

**elpct -0.2184277 0.0269709 -8.099 1.09e-14 \*\*\***

**expnstu 0.0027645 0.0005993 4.613 5.69e-06 \*\*\***

**computer 0.0029127 0.0008671 3.359 0.000874 \*\*\***

mathscr -0.0696774 0.0333254 -2.091 0.037311 \*

compstu -9.0534787 5.6707960 -1.597 0.111335

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.479 on 329 degrees of freedom

Multiple R-squared: 0.6572, Adjusted R-squared: 0.651

F-statistic: 105.1 on 6 and 329 DF, p-value: < 2.2e-16

**5. LASSO**

a. LASSO MIN

14 x 1 sparse Matrix of class "dgCMatrix"

1

(Intercept) 1.941878e-16

.

enrltot .

teachers .

**mealpct 8.628770e-01**

**computer 1.018010e-01**

testscr .

compstu -4.410619e-02

expnstu 1.485089e-01

str .

**avginc -1.362915e-02**

**elpct -3.361761e-01**

readscr .

**mathscr -1.056835e-01**

b. LASSO 1SE

14 x 1 sparse Matrix of class "dgCMatrix"

1

(Intercept) -1.112590e-16

.

enrltot .

teachers .

**mealpct 6.740258e-01**

computer .

testscr .

compstu .

**expnstu 1.981124e-02**

str .

avginc .

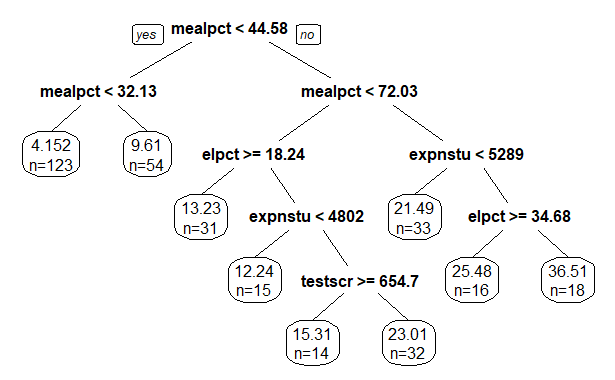
**elpct -2.269198e-02**

readscr .

mathscr .

**6. Regression Tree and Pruned Tree**

a. Regression Tree



When we look at this tree, MealPct is the most influential variable in term of affecting the split of the tree follow by elpct, expnstu, and testscr variable.

Here is the summary of our regression (basic) tree:

1) root 336 40286.0300 12.940580

2) mealpct< 44.57875 177 3200.6190 5.817284

4) mealpct< 32.1306 123 1323.1190 4.152145 \*

5) mealpct>=32.1306 54 759.6450 9.610100 \*

3) mealpct>=44.57875 159 18106.2300 20.870290

6) mealpct< 72.0275 92 4447.5350 16.786040

12) elpct>=18.24072 31 668.3543 13.227160 \*

13) elpct< 18.24072 61 3187.0100 18.594650

26) expnstu< 4801.746 15 279.4655 12.235660 \*

27) expnstu>=4801.746 46 2103.2050 20.668240

54) testscr>=654.675 14 378.7161 15.312870 \*

55) testscr< 654.675 32 1147.3050 23.011210 \*

7) mealpct>=72.0275 67 10016.7400 26.478510

14) expnstu< 5289.162 33 1470.5290 21.492950 \*

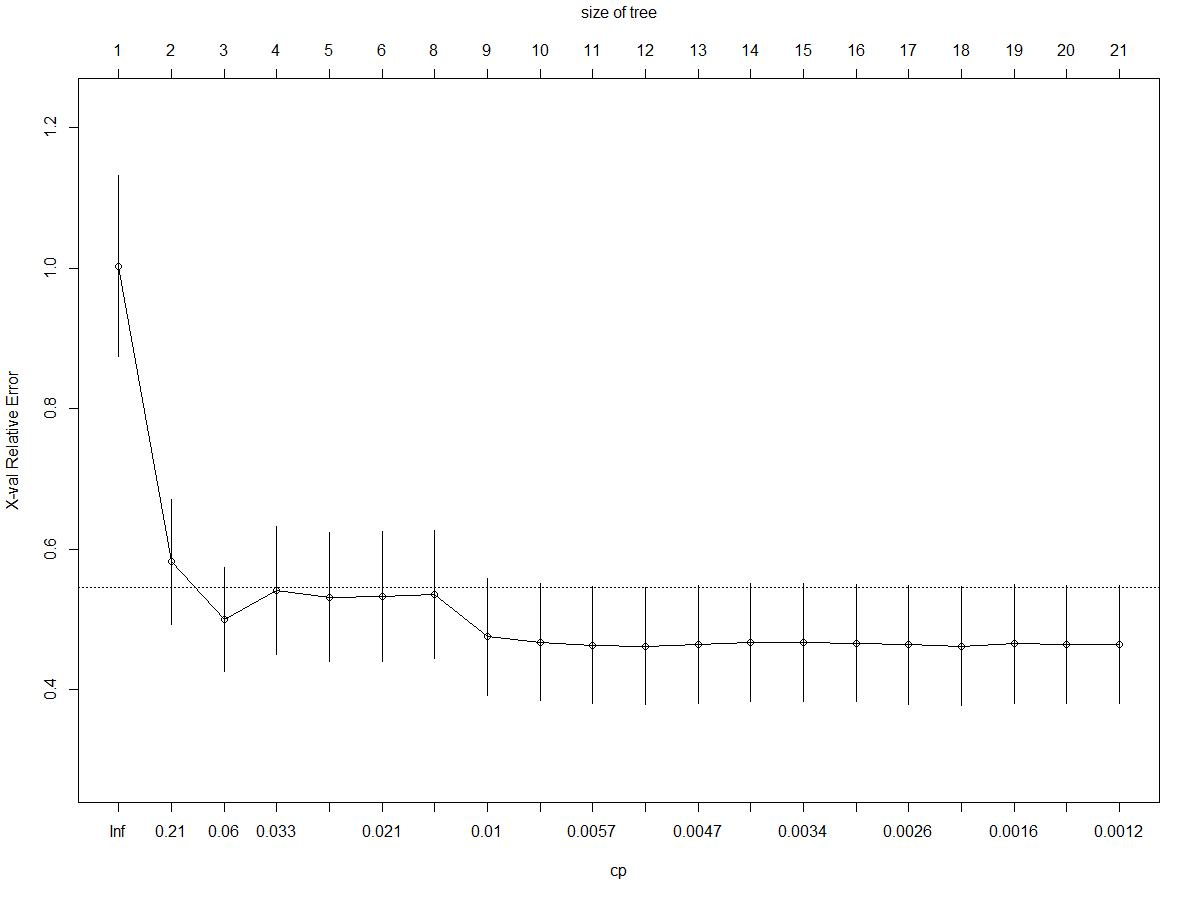
15) expnstu>=5289.162 34 6929.8530 31.317430

30) elpct>=34.67684 16 3668.2180 25.476070 \*

31) elpct< 34.67684 18 2230.4090 36.509750 \*

b. Pruned Tree

After considering the first regression tree, we think we can still make a better tree as we decrease the number of variable selections and increase complexity parameter from 0.001 to 0.0173324 based on the following figures:



With the tree size of 6, we then look up the nsplit=5 in the Displays CP table for Fitted Rpart Object:

Root node error: 40286/336 = 119.9

CP nsplit rel error xerror xstd

1 0.4711108 0 1.00000 1.00334 0.128618

2 0.0904024 1 0.52889 0.58211 0.088372

3 0.0401220 2 0.43849 0.50036 0.074286

4 0.0277480 3 0.39836 0.54088 0.091015

5 0.0255976 4 0.37062 0.53217 0.091289

**6** **0.0173324 5 0.34502 0.53273 0.092060**

7 0.0143271 7 0.31035 0.53527 0.091143

8 0.0074391 8 0.29603 0.47555 0.083632

9 0.0062093 9 0.28859 0.46770 0.083439

10 0.0052107 10 0.28238 0.46341 0.083356

11 0.0052094 11 0.27717 0.46196 0.083353

12 0.0042272 12 0.27196 0.46466 0.083490

13 0.0034058 13 0.26773 0.46703 0.083863

14 0.0034042 14 0.26433 0.46707 0.083909

15 0.0032226 15 0.26092 0.46665 0.083915

16 0.0020732 16 0.25770 0.46388 0.084045

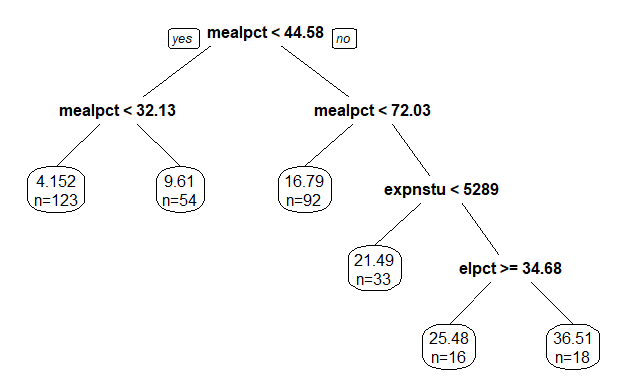
17 0.0017008 17 0.25563 0.46244 0.084051

18 0.0015806 18 0.25393 0.46535 0.084340

19 0.0013656 19 0.25234 0.46438 0.084330

20 0.0010000 20 0.25098 0.46446 0.084121

With the cp=0.0173324 we get the following pruned tree (with 6 nodes):



In this tree, the Mealpct still the most significant variable that affect the Calwork responsive variable following by expnstu and elpct.

**7. Results and Comparison**

a. Results

After carefully building different variable selection models and calculate the related performance measurement of both in-sample and out-of-sample performance. Here is the summary table of the result. In each model, we also list out the top significant explanatory variables that affect the Calwork percent qualify the most.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | ***Model Fitting (In-sample performance)*** | | | | ***Testing Error (out-of-sample performance)*** | |
| **Models** | **Var. name** | **Adj. R^2** | **AIC** | **BIC** | **Model MSE** | **CV score** | **MSPE** |
| **Linear Regression** | elpct, mealpct,  expnstu | 0.6497262 | 2225.107 | 2278.546 | 42.12281 | 44.18613 | 85.84488 |
| **Forward AIC** | Mealpct,  Elpct  Expnstu  Computer | 0.6509847 | 2218.081 | 2248.618 | 41.97147 | 42.99744 | 84.45084 |
| **Backward AIC** | mealpct  expnstu  elpct | 0.6510529 | 2219.967 | 2258.138 | 41.96326 | 43.34812 | 84.56642 |
| **Stepwise AIC** | Mealpct,  Elpct  Expnstu  computer | 0.6509847 | 2218.081 | 2248.618 | 41.97147 | 43.00764 | 84.45084 |
| **LASSO MIN** | Mealpct,  Computer,  Expnstu,  Avginc, elpct, mathscr | 0.6566632 | N/A | N/A | 279.7148 | 47.49423 | 61.17688 |
| **LASSO 1SE** | Meal, expnstu, elpct | 0.5738939 | N/A | N/A | 279.6196 | 56.87361 | 62.59915 |
| **Regression Tree** | Meal,  elpct, testscr,  expnstu | N/A | N/A | N/A | 36.24851 | 54.793101 | 90.79337 |
| **Pruned Tree** | Mealpct, expnstu, elpct | N/A | N/A | N/A | 41.99231 | 63.874327 | 89.56874 |

b. Comparison:

Highest R^2 Square: LASSO Min

Lowest AIC and BIC: Forward AIC and Stepwise AIC

Lowest MSE: Regression Tree

Highest CV-Score: Pruned Tree

Lowest MSPE: LASSO MIN

Top most selected explanatory variables: Mealpct, expnstu, elpct

=> After all consider the above table and conclusion, we choose Pruned Tree as the best model for the Calworks responsive variable in our dataset since it has the highest CV score and fairly low MSE and MSPE.

**8. Conclusion for CalWorks assistance program**

In conclusion, top three factors that affect the percent qualify for Calworks are:

* Percent qualifying for reduced-price lunch (mealpct) has a positive correlation with Calworks because school are more likely to get Calwork assistance program if they are already qualified for Price-Reduced Lunch
* Percent of English Leaner (Elpct) has a negative correlation with Calworks since this program is mainly focusing on US Citizen, the more English learners aka international students in school which mean the less US citizen in that school can result in a lower level of percent qualify for Calworks.
* Expenditure per student (Expnstu): has a positive correlation with Calworks since the more fund a school ask for in order to spend on their students, the more likely they will get more fund for the expenditure per student.
* Interesting Finding: The main reason why Test Score has a negative relationship with Calworks program is because students who eligible for this program are usually those that don’t have the best learning environment which then leads to their poor performances in this case test score. Therefore, student who have lower test score are more likely to get this cash aids so that they can improve their learning environment in order to get better in school performance.

**VI. Variable Selection and Results for Mealpct**

For this part, we run 8 different models in R and also calculate all the associates performance measurement figures for each model. Below is the summary result of model selections for each model. Those highlighted variables are those that is significant in each model.

**1. Linear Regression**

Here is the result we get after running the linear regression.

Call:

lm(formula = mealpct ~ ., data = data.train)

Residuals:

Min 1Q Median 3Q Max

-37.938 -5.886 0.005 5.675 28.746

Coefficients: (1 not defined because of singularities)

Estimate Std. Error t value Pr(>|t|)

(Intercept) 359.953919 39.688207 9.070 < 2e-16 \*\*\*

grspanKK-08 1.967850 1.676317 1.174 0.241295

enrltot -0.000517 0.002113 -0.245 0.806853

teachers 0.011927 0.047162 0.253 0.800513

**calwpct**  0.754006 0.064568 11.678 < 2e-16 \*\*\*

computer -0.003160 0.004066 -0.777 0.437591

testscr 0.100951 0.164457 0.614 0.539750

compstu 17.844253 10.074365 1.771 0.077462 .

expnstu 0.003583 0.001215 2.949 0.003418 \*\*

str 0.118250 0.418071 0.283 0.777475

**avginc** -0.787039 0.124344 -6.330 8.22e-10 \*\*\*

**elpct**  0.359028 0.047681 7.530 5.13e-13 \*\*\*

**readscr** -0.625269 0.163364 -3.827 0.000155 \*\*\*

mathscr NA NA NA NA

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.8 on 323 degrees of freedom

Multiple R-squared: 0.869, Adjusted R-squared: 0.8642

F-statistic: 178.6 on 12 and 323 DF, p-value: < 2.2e-16

**2. Forward Selection**

Call:

lm(formula = mealpct ~ readscr + calwpct + elpct + avginc + expnstu +

teachers + compstu, data = data.train)

Residuals:

Min 1Q Median 3Q Max

-38.720 -5.926 0.084 5.556 29.657

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 3.765e+02 3.613e+01 10.421 < 2e-16 \*\*\*

**readscr** -5.408e-01 5.592e-02 -9.671 < 2e-16 \*\*\*

**calwpct** 7.468e-01 6.341e-02 11.778 < 2e-16 \*\*\*

**elpct** 3.563e-01 4.599e-02 7.746 1.19e-13 \*\*\*

**avginc** -7.824e-01 1.213e-01 -6.449 4.04e-10 \*\*\*

**expnstu** 3.336e-03 9.818e-04 3.398 0.000763 \*\*\*

teachers -5.830e-03 2.977e-03 -1.958 0.051022 .

compstu 1.431e+01 9.030e+00 1.585 0.113966

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.757 on 328 degrees of freedom

Multiple R-squared: 0.8682, Adjusted R-squared: 0.8654

F-statistic: 308.6 on 7 and 328 DF, p-value: < 2.2e-16

**3. Backward Selection**

Call:

lm(formula = mealpct ~ calwpct + computer + compstu + expnstu +

avginc + elpct + readscr, data = data.train)

Residuals:

Min 1Q Median 3Q Max

-39.123 -5.752 0.010 5.559 29.297

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 3.762e+02 3.612e+01 10.416 < 2e-16 \*\*\*

**calwpct** 7.493e-01 6.349e-02 11.801 < 2e-16 \*\*\*

**computer** -2.552e-03 1.247e-03 -2.046 0.04157 \*

compstu 1.699e+01 8.954e+00 1.898 0.05864 .

**expnstu** 3.258e-03 9.864e-04 3.303 0.00106 \*\*

**avginc** -7.752e-01 1.218e-01 -6.364 6.62e-10 \*\*\*

**elpct** 3.555e-01 4.574e-02 7.772 1.00e-13 \*\*\*

**readscr**  -5.404e-01 5.589e-02 -9.668 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.752 on 328 degrees of freedom

Multiple R-squared: 0.8683, Adjusted R-squared: 0.8655

F-statistic: 308.9 on 7 and 328 DF, p-value: < 2.2e-16

**4. Stepwise Selection**

Call:

lm(formula = mealpct ~ readscr + calwpct + elpct + avginc + expnstu +

teachers + compstu, data = data.train)

Residuals:

Min 1Q Median 3Q Max

-38.720 -5.926 0.084 5.556 29.657

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 3.765e+02 3.613e+01 10.421 < 2e-16 \*\*\*

**readscr** -5.408e-01 5.592e-02 -9.671 < 2e-16 \*\*\*

**calwpct** 7.468e-01 6.341e-02 11.778 < 2e-16 \*\*\*

**elpct**  3.563e-01 4.599e-02 7.746 1.19e-13 \*\*\*

**avginc** -7.824e-01 1.213e-01 -6.449 4.04e-10 \*\*\*

**expnstu** 3.336e-03 9.818e-04 3.398 0.000763 \*\*\*

teachers -5.830e-03 2.977e-03 -1.958 0.051022 .

compstu 1.431e+01 9.030e+00 1.585 0.113966

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.757 on 328 degrees of freedom

Multiple R-squared: 0.8682, Adjusted R-squared: 0.8654

F-statistic: 308.6 on 7 and 328 DF, p-value: < 2.2e-16

**5. LASSO**

a. LASSO MIN

14 x 1 sparse Matrix of class "dgCMatrix"

1

(Intercept) -0.099841429

0.053333418

**enrltot -0.001393122**

teachers .

**calwpct 0.318998297**

**computer -0.036245203**

testscr .

**compstu 0.034476285**

**expnstu 0.074894354**

str .

**avginc -0.195925367**

**elpct 0.235437152**

**readscr -0.402707917**

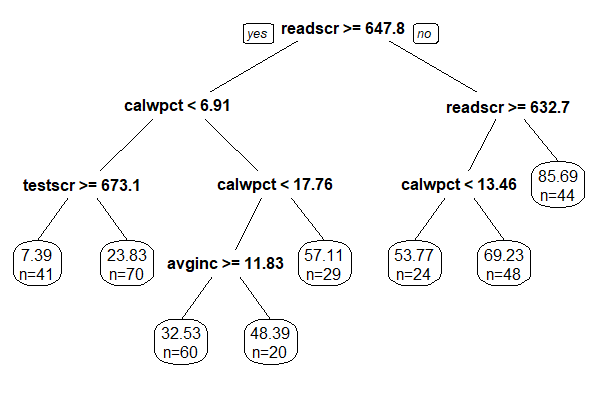
mathscr .

b. LASSO 1SE

|  |
| --- |
| 14 x 1 sparse Matrix of class "dgCMatrix"  1  (Intercept) 2.105258e-16  .  enrltot .  teachers .  **calwpct 2.975088e-01**  computer .  testscr .  compstu .  expnstu .  str .  **avginc -1.222896e-01**  **elpct 1.580847e-01**  **readscr -4.306037e-01**  mathscr . |

**6. Regression Tree and Pruned Tree**

a. Regression Tree



When we look at this tree, Readscr is the most influential variable in term of affecting the split of the tree follow by calwpct, testscr and avginc variable.

Here is the summary of our regression (basic) tree:

1) root 336 236861.200 44.436330

2) readscr>=647.75 220 70476.200 29.757230

4) calwpct< 6.91005 111 15453.190 17.754610

8) testscr>=673.15 41 2689.045 7.390368 \*

9) testscr< 673.15 70 5780.474 23.825100 \*

5) calwpct>=6.91005 109 22747.630 41.980090

10) calwpct< 17.76415 80 10981.380 36.494910

20) avginc>=11.83 60 5690.202 32.529420 \*

21) avginc< 11.83 20 1517.164 48.391350 \*

11) calwpct>=17.76415 29 2719.347 57.111610 \*

3) readscr< 647.75 116 29074.860 72.275990

6) readscr>=632.7 72 11805.700 64.076940

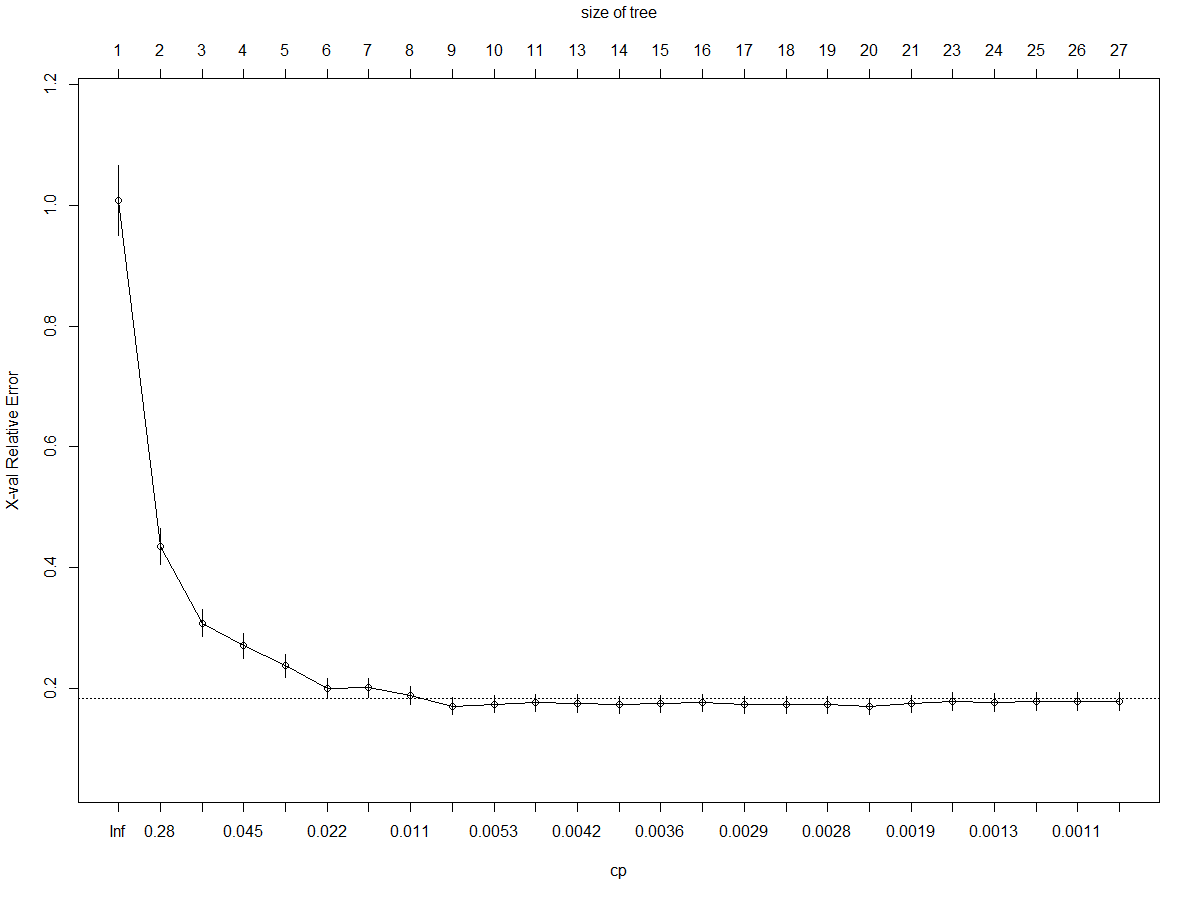
12) calwpct< 13.4629 24 2580.093 53.773390 \*

13) calwpct>=13.4629 48 5403.736 69.228710 \*

7) readscr< 632.7 44 4508.742 85.692620 \*

b. Pruned Tree

After considering the first regression tree, we think we can still make a better tree as we increase complexity parameter from 0.001 to 0.0054622 based on the following figures:



With the tree size of 9, we then look up the nsplit=8 in the Displays CP table for Fitted Rpart Object:

Root node error: 236861/336 = 704.94

CP nsplit rel error xerror xstd

1 0.5797071 0 1.000000 1.00779 0.057592

2 0.1362628 1 0.420293 0.43523 0.030136

3 0.0538730 2 0.284030 0.30792 0.022262

4 0.0381950 3 0.230157 0.27034 0.019998

5 0.0294843 4 0.191962 0.23720 0.018709

6 0.0161355 5 0.162478 0.20061 0.016027

7 0.0159334 6 0.146342 0.20116 0.015658

8 0.0081350 7 0.130409 0.18883 0.015100

**9 0.0054622 8 0.122274 0.17025 0.014114**

10 0.0050650 9 0.116812 0.17392 0.014139

11 0.0042174 10 0.111747 0.17588 0.014007

12 0.0041637 12 0.103312 0.17493 0.014284

13 0.0039179 13 0.099148 0.17254 0.013841

14 0.0033302 14 0.095230 0.17485 0.014179

15 0.0029574 15 0.091900 0.17636 0.014117

16 0.0027996 16 0.088943 0.17304 0.014079

17 0.0027877 17 0.086143 0.17315 0.014157

18 0.0027239 18 0.083356 0.17315 0.014157

19 0.0024847 19 0.080632 0.17012 0.013656

20 0.0014421 20 0.078147 0.17460 0.014266

21 0.0014119 22 0.075263 0.17773 0.014715

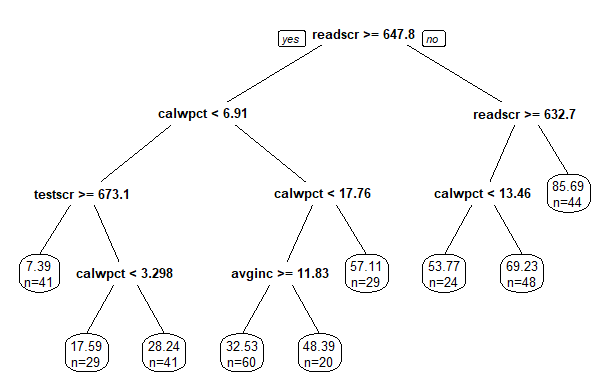
22 0.0012681 23 0.073851 0.17674 0.014621

23 0.0011078 24 0.072583 0.17787 0.014653

24 0.0010370 25 0.071475 0.17779 0.014649

25 0.0010000 26 0.070438 0.17778 0.014650

With the cp=0.0054622 we get the following pruned tree (with 9 nodes):



In this tree, again readscr still the most significant variable that affect the MealPct responsive variable following by calwpct, testscr and avginc.

**7. Results and Comparison**

a. Results

Here is the summary table of the result. In each model, we also list out the top significant explanatory variables that affect the percent qualify for reduced-price lunch (mealpct) the most.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | ***Model Fitting (In-sample performance)*** | | | | ***Testing Error (out-of-sample performance)*** | |
|  | **Var. name** | **Adj. R^2** | **AIC** | **BIC** | **Model MSE** | **CV score** | **MSPE** |
| **Linear Regression** | Calwpct,  expnstu,  avginc,  elpct,  readscr | 0.8641546 | 2502.062 | 2555.501 | 96.04924 | 102.2026 | 109.4401 |
| **Forward AIC** | Readscr, calwpct, elpct, avginc, expnstu, | 0.8653514 | 2494.25 | 2528.604 | 95.20308 | 100.2874 | 112.2624 |
| **Backward AIC** | Calwpct,  computer expnstu, avginc, elpct, readscr | 0.8654932 | 2493.896 | 2528.25 | 95.10283 | 98.74119 | 110.6946 |
| **Stepwise AIC** | Readscr, calwpct, avginc, expnstu  elpct | 0.8653514 | 2494.25 | 2528.604 | 95.20308 | 100.7438 | 112.2624 |
| **LASSO MIN** | Enrltot, calwpct, computer, compstu, expnstu, avginc, elpct, readscr | 0.8686416 | N/A | N/A | 90.5171 | 108.191 | 62.8721 |
| **LASSO 1SE** | Calwpct, avginc, elpct, readscr | 0.8486278 | N/A | N/A | 89.56647 | 119.4584 | 64.0238 |
| **Regression Tree** | Readscr, clwpct, testscr,, avginc | N/A | N/A | N/A | 93.31964 | 129.14008 | 113.6207 |
| **Pruned Tree** | Readscr, calwpct, testscr, avginc | N/A | N/A | N/A | 87.49831 | 133.11382 | 112.9871 |

b. Comparison

Highest R^2 Square: LASSO Min

Lowest AIC and BIC: Backward AIC

Lowest MSE: Pruned Tree

Highest CV-Score: Pruned Tree

Lowest MSPE: LASSO MIN

Top most selected explanatory variables: Calwpct, readscr, testscr and avginc

=> After all consider the above table and comparison, we choose Pruned Tree as the best model for the Mealpct responsive variable in our dataset since it has the highest CV score, lowest MSE and a reasonable MSPE

**8. Conclusion for Price-Reduced Lunch program**

In conclusion, top four factors that affect the percent qualify for Price Reduced Lunch (MealPct) are:

* Percent qualifying for CalWorks (Calwpct) has a positive correlation with Mealpct because school are more likely to get Price-Reduced Lunch assistance program if they are already qualified for CalWorks
* Average Test Score (Testscr) has a negative relationship with Price-Reduced Lunch meaning lowest test score result in higher chance of qualifying for Price-Reduced Lunch because as we discuss above, those student that need these aids are those students whose struggle in school which lead to a lower test score. Therefore, the price-reduced lunch can be considered as an incentive for those student to study and improve their performance.
* Average Income (avginc) has a negative correlation with this program. This is quite easy to understand because lower income household are the one that benefits more from the Price-Reduced Lunch program.
* Interesting Finding: The relationship between average reading score (Readscr) and Mealpct is negative. In our data analysis, this make sense because Average Test Score has a negative correlation with Mealpct and Average Test Score = (Read + Math)/2, this lead to the Readscr is also negatively correlate with Mealpct. However, in real-life concept, this is quite an interesting finding between Average Reading Score and Percent Qualify for Reduced- Price Lunch