**Decision Tree Classification**

**with Titanic dataset**

**For the train dataset**

**preprocessing of the dataset and exploration about the data**

1. check the table using view train, found out some columns are useful to predict the survived rate, so I Only choose Pclass + Sex + Age + SibSp + Parch + Fare + Embarked.
2. Did the data cleaning, check rows and columns, if they are empty, and replace with na.
3. Check the absolute rate of survived to have around percentage for exploring data.
4. Check the coefficient between the survived rate with other factors, and see how it got effected.

**During the process of modeling:**

1. Build classification decision tree using rpart, save the result in fit.
2. Before building the tree model, we need to balance two aspects, the difference after classification should be smaller, and also need to avoid matching too much and make the error od model becomes larger. so we need to consider CP and Xerror, to make CP smaller when Xerror is smallest.
3. First use printcp got the all kinds of error, the Root node error and Xerror
4. Apply plotcp graphic about CP and Xerror, find the CP value when Xerror smallest, decide the CP value
5. Cut some branches which are not effective using prune, Prune the dataset and plot prune tree , and create the ps file
6. The reason for every steps are as follows below:

> # Classification Tree with rpart

> #install.packages("party")

> library(rpart)

> test<-read.csv("C:/Users/Wanwan Zhang/Desktop/2016FALL/ADS/5/test.csv")

> train<-read.csv("C:/Users/Wanwan Zhang/Desktop/2016FALL/ADS/5/train.csv")

>

> str(test)

'data.frame': 418 obs. of 11 variables:

$ PassengerId: int 892 893 894 895 896 897 898 899 900 901 ...

$ Pclass : int 3 3 2 3 3 3 3 2 3 3 ...

$ Name : Factor w/ 418 levels "Abbott, Master. Eugene Joseph",..: 210 409 273 414 182 370 85 58 5 104 ...

$ Sex : Factor w/ 2 levels "female","male": 2 1 2 2 1 2 1 2 1 2 ...

$ Age : num 34.5 47 62 27 22 14 30 26 18 21 ...

$ SibSp : int 0 1 0 0 1 0 0 1 0 2 ...

$ Parch : int 0 0 0 0 1 0 0 1 0 0 ...

$ Ticket : Factor w/ 363 levels "110469","110489",..: 153 222 74 148 139 262 159 85 101 270 ...

$ Fare : num 7.83 7 9.69 8.66 12.29 ...

$ Cabin : Factor w/ 77 levels "","A11","A18",..: 1 1 1 1 1 1 1 1 1 1 ...

$ Embarked : Factor w/ 3 levels "C","Q","S": 2 3 2 3 3 3 2 3 1 3 ...

> str(train)

'data.frame': 891 obs. of 12 variables:

$ PassengerId: int 1 2 3 4 5 6 7 8 9 10 ...

$ Survived : int 0 1 1 1 0 0 0 0 1 1 ...

$ Pclass : int 3 1 3 1 3 3 1 3 3 2 ...

$ Name : Factor w/ 891 levels "Abbing, Mr. Anthony",..: 109 191 358 277 16 559 520 629 417 581 ...

$ Sex : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 2 1 1 ...

$ Age : num 22 38 26 35 35 NA 54 2 27 14 ...

$ SibSp : int 1 1 0 1 0 0 0 3 0 1 ...

$ Parch : int 0 0 0 0 0 0 0 1 2 0 ...

$ Ticket : Factor w/ 681 levels "110152","110413",..: 524 597 670 50 473 276 86 396 345 133 ...

$ Fare : num 7.25 71.28 7.92 53.1 8.05 ...

$ Cabin : Factor w/ 148 levels "","A10","A14",..: 1 83 1 57 1 1 131 1 1 1 ...

$ Embarked : Factor w/ 4 levels "","C","Q","S": 4 2 4 4 4 3 4 4 4 2 ...

> # Check Survival rates of all people

> table(train$Survived)

0 1

549 342

>

> # Survival rates in proportions

> prop.table(table(train$Survived))

0 1

0.6161616 0.3838384

> train1 <- train

> ##list of rows with missing values

> train1[!complete.cases(train1),]

PassengerId Survived Pclass Name Sex Age SibSp

6 6 0 3 Moran, Mr. James male NA 0

18 18 1 2 Williams, Mr. Charles Eugene male NA 0

20 20 1 3 Masselmani, Mrs. Fatima female NA 0

27 27 0 3 Emir, Mr. Farred Chehab male NA 0

29 29 1 3 O'Dwyer, Miss. Ellen "Nellie" female NA 0

30 30 0 3 Todoroff, Mr. Lalio male NA 0

32 32 1 1 Spencer, Mrs. William Augustus (Marie Eugenie) female NA 1

33 33 1 3 Glynn, Miss. Mary Agatha female NA 0

37 37 1 3 Mamee, Mr. Hanna male NA 0

43 43 0 3 Kraeff, Mr. Theodor male NA 0

46 46 0 3 Rogers, Mr. William John male NA 0

47 47 0 3 Lennon, Mr. Denis male NA 1

48 48 1 3 O'Driscoll, Miss. Bridget female NA 0

49 49 0 3 Samaan, Mr. Youssef male NA 2

56 56 1 1 Woolner, Mr. Hugh male NA 0

65 65 0 1 Stewart, Mr. Albert A male NA 0

66 66 1 3 Moubarek, Master. Gerios male NA 1

77 77 0 3 Staneff, Mr. Ivan male NA 0

78 78 0 3 Moutal, Mr. Rahamin Haim male NA 0

83 83 1 3 McDermott, Miss. Brigdet Delia female NA 0

88 88 0 3 Slocovski, Mr. Selman Francis male NA 0

96 96 0 3 Shorney, Mr. Charles Joseph male NA 0

102 102 0 3 Petroff, Mr. Pastcho ("Pentcho") male NA 0

108 108 1 3 Moss, Mr. Albert Johan male NA 0

110 110 1 3 Moran, Miss. Bertha female NA 1

122 122 0 3 Moore, Mr. Leonard Charles male NA 0

127 127 0 3 McMahon, Mr. Martin male NA 0

129 129 1 3 Peter, Miss. Anna female NA 1

141 141 0 3 Boulos, Mrs. Joseph (Sultana) female NA 0

155 155 0 3 Olsen, Mr. Ole Martin male NA 0

159 159 0 3 Smiljanic, Mr. Mile male NA 0

160 160 0 3 Sage, Master. Thomas Henry male NA 8

167 167 1 1 Chibnall, Mrs. (Edith Martha Bowerman) female NA 0

169 169 0 1 Baumann, Mr. John D male NA 0

177 177 0 3 Lefebre, Master. Henry Forbes male NA 3

181 181 0 3 Sage, Miss. Constance Gladys female NA 8

182 182 0 2 Pernot, Mr. Rene male NA 0

186 186 0 1 Rood, Mr. Hugh Roscoe male NA 0

187 187 1 3 O'Brien, Mrs. Thomas (Johanna "Hannah" Godfrey) female NA 1

197 197 0 3 Mernagh, Mr. Robert male NA 0

199 199 1 3 Madigan, Miss. Margaret "Maggie" female NA 0

202 202 0 3 Sage, Mr. Frederick male NA 8

215 215 0 3 Kiernan, Mr. Philip male NA 1

224 224 0 3 Nenkoff, Mr. Christo male NA 0

230 230 0 3 Lefebre, Miss. Mathilde female NA 3

236 236 0 3 Harknett, Miss. Alice Phoebe female NA 0

241 241 0 3 Zabour, Miss. Thamine female NA 1

242 242 1 3 Murphy, Miss. Katherine "Kate" female NA 1

251 251 0 3 Reed, Mr. James George male NA 0

257 257 1 1 Thorne, Mrs. Gertrude Maybelle female NA 0

261 261 0 3 Smith, Mr. Thomas male NA 0

265 265 0 3 Henry, Miss. Delia female NA 0

271 271 0 1 Cairns, Mr. Alexander male NA 0

275 275 1 3 Healy, Miss. Hanora "Nora" female NA 0

278 278 0 2 Parkes, Mr. Francis "Frank" male NA 0

285 285 0 1 Smith, Mr. Richard William male NA 0

296 296 0 1 Lewy, Mr. Ervin G male NA 0

299 299 1 1 Saalfeld, Mr. Adolphe male NA 0

301 301 1 3 Kelly, Miss. Anna Katherine "Annie Kate" female NA 0

302 302 1 3 McCoy, Mr. Bernard male NA 2

304 304 1 2 Keane, Miss. Nora A female NA 0

305 305 0 3 Williams, Mr. Howard Hugh "Harry" male NA 0

307 307 1 1 Fleming, Miss. Margaret female NA 0

325 325 0 3 Sage, Mr. George John Jr male NA 8

331 331 1 3 McCoy, Miss. Agnes female NA 2

335 335 1 1 Frauenthal, Mrs. Henry William (Clara Heinsheimer) female NA 1

336 336 0 3 Denkoff, Mr. Mitto male NA 0

348 348 1 3 Davison, Mrs. Thomas Henry (Mary E Finck) female NA 1

352 352 0 1 Williams-Lambert, Mr. Fletcher Fellows male NA 0

355 355 0 3 Yousif, Mr. Wazli male NA 0

359 359 1 3 McGovern, Miss. Mary female NA 0

360 360 1 3 Mockler, Miss. Helen Mary "Ellie" female NA 0

365 365 0 3 O'Brien, Mr. Thomas male NA 1

368 368 1 3 Moussa, Mrs. (Mantoura Boulos) female NA 0

369 369 1 3 Jermyn, Miss. Annie female NA 0

376 376 1 1 Meyer, Mrs. Edgar Joseph (Leila Saks) female NA 1

385 385 0 3 Plotcharsky, Mr. Vasil male NA 0

389 389 0 3 Sadlier, Mr. Matthew male NA 0

410 410 0 3 Lefebre, Miss. Ida female NA 3

411 411 0 3 Sdycoff, Mr. Todor male NA 0

412 412 0 3 Hart, Mr. Henry male NA 0

414 414 0 2 Cunningham, Mr. Alfred Fleming male NA 0

416 416 0 3 Meek, Mrs. Thomas (Annie Louise Rowley) female NA 0

421 421 0 3 Gheorgheff, Mr. Stanio male NA 0

426 426 0 3 Wiseman, Mr. Phillippe male NA 0

429 429 0 3 Flynn, Mr. James male NA 0

432 432 1 3 Thorneycroft, Mrs. Percival (Florence Kate White) female NA 1

445 445 1 3 Johannesen-Bratthammer, Mr. Bernt male NA 0

452 452 0 3 Hagland, Mr. Ingvald Olai Olsen male NA 1

455 455 0 3 Peduzzi, Mr. Joseph male NA 0

458 458 1 1 Kenyon, Mrs. Frederick R (Marion) female NA 1

460 460 0 3 O'Connor, Mr. Maurice male NA 0

465 465 0 3 Maisner, Mr. Simon male NA 0

467 467 0 2 Campbell, Mr. William male NA 0

469 469 0 3 Scanlan, Mr. James male NA 0

471 471 0 3 Keefe, Mr. Arthur male NA 0

476 476 0 1 Clifford, Mr. George Quincy male NA 0

482 482 0 2 Frost, Mr. Anthony Wood "Archie" male NA 0

486 486 0 3 Lefebre, Miss. Jeannie female NA 3

491 491 0 3 Hagland, Mr. Konrad Mathias Reiersen male NA 1

496 496 0 3 Yousseff, Mr. Gerious male NA 0

498 498 0 3 Shellard, Mr. Frederick William male NA 0

503 503 0 3 O'Sullivan, Miss. Bridget Mary female NA 0

508 508 1 1 Bradley, Mr. George ("George Arthur Brayton") male NA 0

512 512 0 3 Webber, Mr. James male NA 0

518 518 0 3 Ryan, Mr. Patrick male NA 0

523 523 0 3 Lahoud, Mr. Sarkis male NA 0

525 525 0 3 Kassem, Mr. Fared male NA 0

528 528 0 1 Farthing, Mr. John male NA 0

532 532 0 3 Toufik, Mr. Nakli male NA 0

534 534 1 3 Peter, Mrs. Catherine (Catherine Rizk) female NA 0

539 539 0 3 Risien, Mr. Samuel Beard male NA 0

548 548 1 2 Padro y Manent, Mr. Julian male NA 0

553 553 0 3 O'Brien, Mr. Timothy male NA 0

558 558 0 1 Robbins, Mr. Victor male NA 0

561 561 0 3 Morrow, Mr. Thomas Rowan male NA 0

564 564 0 3 Simmons, Mr. John male NA 0

565 565 0 3 Meanwell, Miss. (Marion Ogden) female NA 0

569 569 0 3 Doharr, Mr. Tannous male NA 0

574 574 1 3 Kelly, Miss. Mary female NA 0

579 579 0 3 Caram, Mrs. Joseph (Maria Elias) female NA 1

585 585 0 3 Paulner, Mr. Uscher male NA 0

590 590 0 3 Murdlin, Mr. Joseph male NA 0

594 594 0 3 Bourke, Miss. Mary female NA 0

597 597 1 2 Leitch, Miss. Jessie Wills female NA 0

599 599 0 3 Boulos, Mr. Hanna male NA 0

602 602 0 3 Slabenoff, Mr. Petco male NA 0

603 603 0 1 Harrington, Mr. Charles H male NA 0

612 612 0 3 Jardin, Mr. Jose Neto male NA 0

613 613 1 3 Murphy, Miss. Margaret Jane female NA 1

614 614 0 3 Horgan, Mr. John male NA 0

630 630 0 3 O'Connell, Mr. Patrick D male NA 0

634 634 0 1 Parr, Mr. William Henry Marsh male NA 0

640 640 0 3 Thorneycroft, Mr. Percival male NA 1

644 644 1 3 Foo, Mr. Choong male NA 0

649 649 0 3 Willey, Mr. Edward male NA 0

651 651 0 3 Mitkoff, Mr. Mito male NA 0

654 654 1 3 O'Leary, Miss. Hanora "Norah" female NA 0

657 657 0 3 Radeff, Mr. Alexander male NA 0

668 668 0 3 Rommetvedt, Mr. Knud Paust male NA 0

670 670 1 1 Taylor, Mrs. Elmer Zebley (Juliet Cummins Wright) female NA 1

675 675 0 2 Watson, Mr. Ennis Hastings male NA 0

681 681 0 3 Peters, Miss. Katie female NA 0

693 693 1 3 Lam, Mr. Ali male NA 0

698 698 1 3 Mullens, Miss. Katherine "Katie" female NA 0

710 710 1 3 Moubarek, Master. Halim Gonios ("William George") male NA 1

712 712 0 1 Klaber, Mr. Herman male NA 0

719 719 0 3 McEvoy, Mr. Michael male NA 0

728 728 1 3 Mannion, Miss. Margareth female NA 0

733 733 0 2 Knight, Mr. Robert J male NA 0

739 739 0 3 Ivanoff, Mr. Kanio male NA 0

740 740 0 3 Nankoff, Mr. Minko male NA 0

741 741 1 1 Hawksford, Mr. Walter James male NA 0

761 761 0 3 Garfirth, Mr. John male NA 0

767 767 0 1 Brewe, Dr. Arthur Jackson male NA 0

769 769 0 3 Moran, Mr. Daniel J male NA 1

774 774 0 3 Elias, Mr. Dibo male NA 0

777 777 0 3 Tobin, Mr. Roger male NA 0

779 779 0 3 Kilgannon, Mr. Thomas J male NA 0

784 784 0 3 Johnston, Mr. Andrew G male NA 1

791 791 0 3 Keane, Mr. Andrew "Andy" male NA 0

793 793 0 3 Sage, Miss. Stella Anna female NA 8

794 794 0 1 Hoyt, Mr. William Fisher male NA 0

816 816 0 1 Fry, Mr. Richard male NA 0

826 826 0 3 Flynn, Mr. John male NA 0

827 827 0 3 Lam, Mr. Len male NA 0

829 829 1 3 McCormack, Mr. Thomas Joseph male NA 0

833 833 0 3 Saad, Mr. Amin male NA 0

838 838 0 3 Sirota, Mr. Maurice male NA 0

840 840 1 1 Marechal, Mr. Pierre male NA 0

847 847 0 3 Sage, Mr. Douglas Bullen male NA 8

850 850 1 1 Goldenberg, Mrs. Samuel L (Edwiga Grabowska) female NA 1

860 860 0 3 Razi, Mr. Raihed male NA 0

864 864 0 3 Sage, Miss. Dorothy Edith "Dolly" female NA 8

869 869 0 3 van Melkebeke, Mr. Philemon male NA 0

879 879 0 3 Laleff, Mr. Kristo male NA 0

889 889 0 3 Johnston, Miss. Catherine Helen "Carrie" female NA 1

Parch Ticket Fare Cabin Embarked

6 0 330877 8.4583 Q

18 0 244373 13.0000 S

20 0 2649 7.2250 C

27 0 2631 7.2250 C

29 0 330959 7.8792 Q

30 0 349216 7.8958 S

32 0 PC 17569 146.5208 B78 C

33 0 335677 7.7500 Q

37 0 2677 7.2292 C

43 0 349253 7.8958 C

46 0 S.C./A.4. 23567 8.0500 S

47 0 370371 15.5000 Q

48 0 14311 7.7500 Q

49 0 2662 21.6792 C

56 0 19947 35.5000 C52 S

65 0 PC 17605 27.7208 C

66 1 2661 15.2458 C

77 0 349208 7.8958 S

78 0 374746 8.0500 S

83 0 330932 7.7875 Q

88 0 SOTON/OQ 392086 8.0500 S

96 0 374910 8.0500 S

102 0 349215 7.8958 S

108 0 312991 7.7750 S

110 0 371110 24.1500 Q

122 0 A4. 54510 8.0500 S

127 0 370372 7.7500 Q

129 1 2668 22.3583 F E69 C

141 2 2678 15.2458 C

155 0 Fa 265302 7.3125 S

159 0 315037 8.6625 S

160 2 CA. 2343 69.5500 S

167 1 113505 55.0000 E33 S

169 0 PC 17318 25.9250 S

177 1 4133 25.4667 S

181 2 CA. 2343 69.5500 S

182 0 SC/PARIS 2131 15.0500 C

186 0 113767 50.0000 A32 S

187 0 370365 15.5000 Q

197 0 368703 7.7500 Q

199 0 370370 7.7500 Q

202 2 CA. 2343 69.5500 S

215 0 367229 7.7500 Q

224 0 349234 7.8958 S

230 1 4133 25.4667 S

236 0 W./C. 6609 7.5500 S

241 0 2665 14.4542 C

242 0 367230 15.5000 Q

251 0 362316 7.2500 S

257 0 PC 17585 79.2000 C

261 0 384461 7.7500 Q

265 0 382649 7.7500 Q

271 0 113798 31.0000 S

275 0 370375 7.7500 Q

278 0 239853 0.0000 S

285 0 113056 26.0000 A19 S

296 0 PC 17612 27.7208 C

299 0 19988 30.5000 C106 S

301 0 9234 7.7500 Q

302 0 367226 23.2500 Q

304 0 226593 12.3500 E101 Q

305 0 A/5 2466 8.0500 S

307 0 17421 110.8833 C

325 2 CA. 2343 69.5500 S

331 0 367226 23.2500 Q

335 0 PC 17611 133.6500 S

336 0 349225 7.8958 S

348 0 386525 16.1000 S

352 0 113510 35.0000 C128 S

355 0 2647 7.2250 C

359 0 330931 7.8792 Q

360 0 330980 7.8792 Q

365 0 370365 15.5000 Q

368 0 2626 7.2292 C

369 0 14313 7.7500 Q

376 0 PC 17604 82.1708 C

385 0 349227 7.8958 S

389 0 367655 7.7292 Q

410 1 4133 25.4667 S

411 0 349222 7.8958 S

412 0 394140 6.8583 Q

414 0 239853 0.0000 S

416 0 343095 8.0500 S

421 0 349254 7.8958 C

426 0 A/4. 34244 7.2500 S

429 0 364851 7.7500 Q

432 0 376564 16.1000 S

445 0 65306 8.1125 S

452 0 65303 19.9667 S

455 0 A/5 2817 8.0500 S

458 0 17464 51.8625 D21 S

460 0 371060 7.7500 Q

465 0 A/S 2816 8.0500 S

467 0 239853 0.0000 S

469 0 36209 7.7250 Q

471 0 323592 7.2500 S

476 0 110465 52.0000 A14 S

482 0 239854 0.0000 S

486 1 4133 25.4667 S

491 0 65304 19.9667 S

496 0 2627 14.4583 C

498 0 C.A. 6212 15.1000 S

503 0 330909 7.6292 Q

508 0 111427 26.5500 S

512 0 SOTON/OQ 3101316 8.0500 S

518 0 371110 24.1500 Q

523 0 2624 7.2250 C

525 0 2700 7.2292 C

528 0 PC 17483 221.7792 C95 S

532 0 2641 7.2292 C

534 2 2668 22.3583 C

539 0 364498 14.5000 S

548 0 SC/PARIS 2146 13.8625 C

553 0 330979 7.8292 Q

558 0 PC 17757 227.5250 C

561 0 372622 7.7500 Q

564 0 SOTON/OQ 392082 8.0500 S

565 0 SOTON/O.Q. 392087 8.0500 S

569 0 2686 7.2292 C

574 0 14312 7.7500 Q

579 0 2689 14.4583 C

585 0 3411 8.7125 C

590 0 A./5. 3235 8.0500 S

594 2 364848 7.7500 Q

597 0 248727 33.0000 S

599 0 2664 7.2250 C

602 0 349214 7.8958 S

603 0 113796 42.4000 S

612 0 SOTON/O.Q. 3101305 7.0500 S

613 0 367230 15.5000 Q

614 0 370377 7.7500 Q

630 0 334912 7.7333 Q

634 0 112052 0.0000 S

640 0 376564 16.1000 S

644 0 1601 56.4958 S

649 0 S.O./P.P. 751 7.5500 S

651 0 349221 7.8958 S

654 0 330919 7.8292 Q

657 0 349223 7.8958 S

668 0 312993 7.7750 S

670 0 19996 52.0000 C126 S

675 0 239856 0.0000 S

681 0 330935 8.1375 Q

693 0 1601 56.4958 S

698 0 35852 7.7333 Q

710 1 2661 15.2458 C

712 0 113028 26.5500 C124 S

719 0 36568 15.5000 Q

728 0 36866 7.7375 Q

733 0 239855 0.0000 S

739 0 349201 7.8958 S

740 0 349218 7.8958 S

741 0 16988 30.0000 D45 S

761 0 358585 14.5000 S

767 0 112379 39.6000 C

769 0 371110 24.1500 Q

774 0 2674 7.2250 C

777 0 383121 7.7500 F38 Q

779 0 36865 7.7375 Q

784 2 W./C. 6607 23.4500 S

791 0 12460 7.7500 Q

793 2 CA. 2343 69.5500 S

794 0 PC 17600 30.6958 C

816 0 112058 0.0000 B102 S

826 0 368323 6.9500 Q

827 0 1601 56.4958 S

829 0 367228 7.7500 Q

833 0 2671 7.2292 C

838 0 392092 8.0500 S

840 0 11774 29.7000 C47 C

847 2 CA. 2343 69.5500 S

850 0 17453 89.1042 C92 C

860 0 2629 7.2292 C

864 2 CA. 2343 69.5500 S

869 0 345777 9.5000 S

879 0 349217 7.8958 S

889 2 W./C. 6607 23.4500 S

> ##list of columns with missing values

> ##if any missing values are there omit them

> train1<-na.omit(train1,na.action = TRUE)

> # grow tree

> fit<- rpart(Survived ~ Pclass + Sex + Age + SibSp + Parch + Fare + Embarked, data = train1, method = "class")

> printcp(fit)# results

Classification tree:

rpart(formula = Survived ~ Pclass + Sex + Age + SibSp + Parch +

Fare + Embarked, data = train1, method = "class")

Variables actually used in tree construction:

[1] Age Fare Parch Pclass Sex SibSp

Root node error: 290/714 = 0.40616

n= 714

CP nsplit rel error xerror xstd

1 0.458621 0 1.00000 1.00000 0.045252

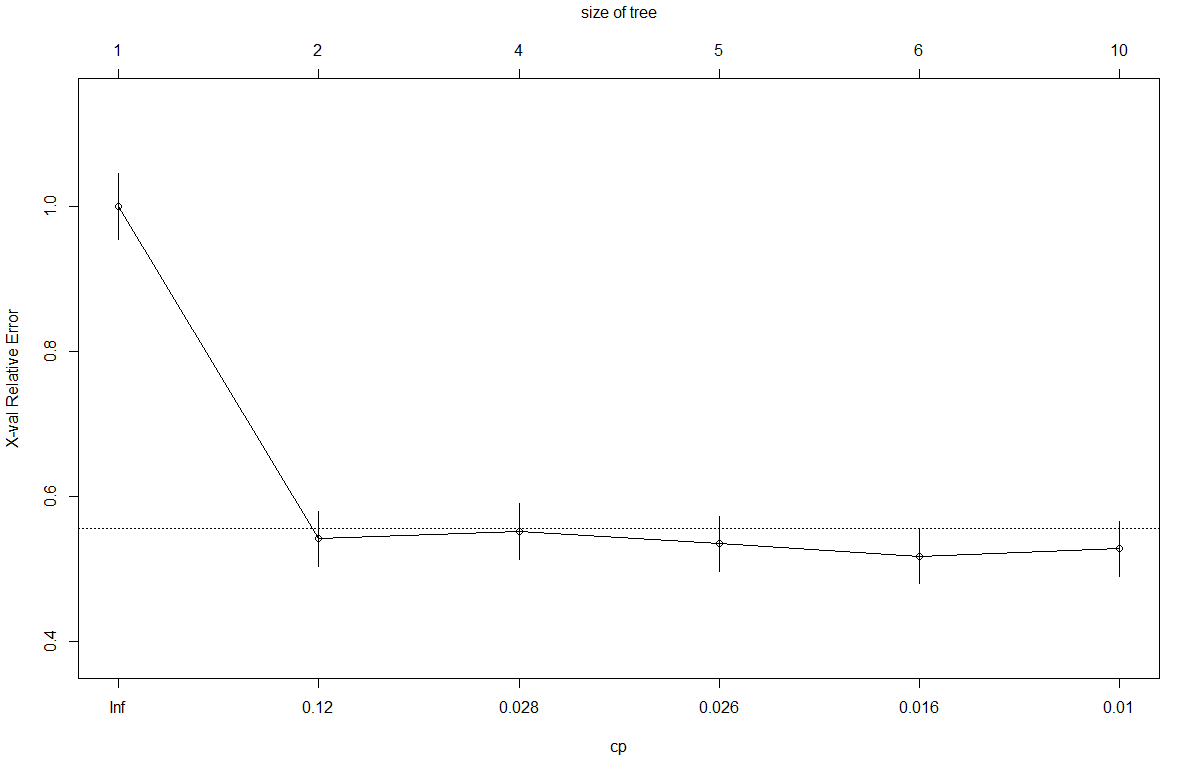
2 0.029310 1 0.54138 0.54138 0.038162

3 0.027586 3 0.48276 0.55172 0.038421

4 0.024138 4 0.45517 0.53448 0.037986

5 0.010345 5 0.43103 0.51724 0.037535

6 0.010000 9 0.38966 0.52759 0.037808



> plotcp(fit)# visualize cross-validation results

> summary(fit)#detailed summary of splits

Call:

rpart(formula = Survived ~ Pclass + Sex + Age + SibSp + Parch +

Fare + Embarked, data = train1, method = "class")

n= 714

CP nsplit rel error xerror xstd

1 0.45862069 0 1.0000000 1.0000000 0.04525169

2 0.02931034 1 0.5413793 0.5413793 0.03816193

3 0.02758621 3 0.4827586 0.5517241 0.03842092

4 0.02413793 4 0.4551724 0.5344828 0.03798609

5 0.01034483 5 0.4310345 0.5172414 0.03753514

6 0.01000000 9 0.3896552 0.5275862 0.03780768

Variable importance

Sex Fare Pclass Age Parch SibSp Embarked

44 15 15 10 7 6 3

Node number 1: 714 observations, complexity param=0.4586207

predicted class=0 expected loss=0.4061625 P(node) =1

class counts: 424 290

probabilities: 0.594 0.406

left son=2 (453 obs) right son=3 (261 obs)

Primary splits:

Sex splits as RL, improve=99.998170, (0 missing)

Pclass < 2.5 to the right, improve=39.252560, (0 missing)

Fare < 52.2771 to the left, improve=33.494530, (0 missing)

Embarked splits as RRLL, improve=13.941390, (0 missing)

Parch < 0.5 to the left, improve= 9.314814, (0 missing)

Surrogate splits:

Fare < 77.6229 to the left, agree=0.668, adj=0.092, (0 split)

Parch < 0.5 to the left, agree=0.667, adj=0.088, (0 split)

Age < 15.5 to the right, agree=0.639, adj=0.011, (0 split)

Embarked splits as RLLL, agree=0.637, adj=0.008, (0 split)

Node number 2: 453 observations, complexity param=0.02758621

predicted class=0 expected loss=0.205298 P(node) =0.6344538

class counts: 360 93

probabilities: 0.795 0.205

left son=4 (429 obs) right son=5 (24 obs)

Primary splits:

Age < 6.5 to the right, improve=10.788930, (0 missing)

Pclass < 1.5 to the right, improve= 9.457965, (0 missing)

Fare < 26.26875 to the left, improve= 8.063017, (0 missing)

Embarked splits as -RLL, improve= 3.307097, (0 missing)

Parch < 0.5 to the left, improve= 2.641424, (0 missing)

Node number 3: 261 observations, complexity param=0.02931034

predicted class=1 expected loss=0.2452107 P(node) =0.3655462

class counts: 64 197

probabilities: 0.245 0.755

left son=6 (102 obs) right son=7 (159 obs)

Primary splits:

Pclass < 2.5 to the right, improve=28.945620, (0 missing)

Fare < 48.2 to the left, improve=10.499530, (0 missing)

Parch < 3.5 to the right, improve= 5.386930, (0 missing)

SibSp < 2.5 to the right, improve= 4.006523, (0 missing)

Embarked splits as RRLL, improve= 3.735682, (0 missing)

Surrogate splits:

Fare < 22.5125 to the left, agree=0.785, adj=0.451, (0 split)

Age < 18.5 to the left, agree=0.678, adj=0.176, (0 split)

SibSp < 1.5 to the right, agree=0.644, adj=0.088, (0 split)

Embarked splits as RRLR, agree=0.640, adj=0.078, (0 split)

Parch < 2.5 to the right, agree=0.636, adj=0.069, (0 split)

Node number 4: 429 observations

predicted class=0 expected loss=0.1794872 P(node) =0.6008403

class counts: 352 77

probabilities: 0.821 0.179

Node number 5: 24 observations, complexity param=0.02413793

predicted class=1 expected loss=0.3333333 P(node) =0.03361345

class counts: 8 16

probabilities: 0.333 0.667

left son=10 (9 obs) right son=11 (15 obs)

Primary splits:

SibSp < 2.5 to the right, improve=8.8888890, (0 missing)

Pclass < 2.5 to the right, improve=3.8095240, (0 missing)

Fare < 20.825 to the right, improve=2.6666670, (0 missing)

Age < 1.5 to the right, improve=0.6095238, (0 missing)

Surrogate splits:

Pclass < 2.5 to the right, agree=0.792, adj=0.444, (0 split)

Fare < 26.95 to the right, agree=0.750, adj=0.333, (0 split)

Embarked splits as -RLR, agree=0.708, adj=0.222, (0 split)

Node number 6: 102 observations, complexity param=0.02931034

predicted class=0 expected loss=0.4607843 P(node) =0.1428571

class counts: 55 47

probabilities: 0.539 0.461

left son=12 (23 obs) right son=13 (79 obs)

Primary splits:

Fare < 20.8 to the right, improve=6.481541, (0 missing)

Age < 38.5 to the right, improve=3.875163, (0 missing)

SibSp < 0.5 to the right, improve=2.138138, (0 missing)

Embarked splits as -RLL, improve=1.950809, (0 missing)

Parch < 3.5 to the right, improve=1.519357, (0 missing)

Surrogate splits:

Parch < 1.5 to the right, agree=0.902, adj=0.565, (0 split)

SibSp < 2.5 to the right, agree=0.853, adj=0.348, (0 split)

Age < 12 to the left, agree=0.784, adj=0.043, (0 split)

Node number 7: 159 observations

predicted class=1 expected loss=0.05660377 P(node) =0.2226891

class counts: 9 150

probabilities: 0.057 0.943

Node number 10: 9 observations

predicted class=0 expected loss=0.1111111 P(node) =0.01260504

class counts: 8 1

probabilities: 0.889 0.111

Node number 11: 15 observations

predicted class=1 expected loss=0 P(node) =0.0210084

class counts: 0 15

probabilities: 0.000 1.000

Node number 12: 23 observations

predicted class=0 expected loss=0.1304348 P(node) =0.03221289

class counts: 20 3

probabilities: 0.870 0.130

Node number 13: 79 observations, complexity param=0.01034483

predicted class=1 expected loss=0.443038 P(node) =0.1106443

class counts: 35 44

probabilities: 0.443 0.557

left son=26 (59 obs) right son=27 (20 obs)

Primary splits:

Age < 16.5 to the right, improve=3.1636130, (0 missing)

Fare < 7.74165 to the right, improve=1.8007220, (0 missing)

Parch < 1.5 to the left, improve=1.3841670, (0 missing)

Embarked splits as -RLL, improve=0.6837703, (0 missing)

SibSp < 0.5 to the right, improve=0.2814594, (0 missing)

Surrogate splits:

Embarked splits as -RLL, agree=0.823, adj=0.30, (0 split)

Fare < 7.2396 to the right, agree=0.759, adj=0.05, (0 split)

Node number 26: 59 observations, complexity param=0.01034483

predicted class=0 expected loss=0.4745763 P(node) =0.08263305

class counts: 31 28

probabilities: 0.525 0.475

left son=52 (7 obs) right son=53 (52 obs)

Primary splits:

Age < 36.5 to the right, improve=1.7479050, (0 missing)

Parch < 0.5 to the left, improve=0.6328197, (0 missing)

SibSp < 0.5 to the right, improve=0.6316236, (0 missing)

Fare < 8.29375 to the right, improve=0.6193810, (0 missing)

Embarked splits as -LLR, improve=0.3328197, (0 missing)

Node number 27: 20 observations

predicted class=1 expected loss=0.2 P(node) =0.0280112

class counts: 4 16

probabilities: 0.200 0.800

Node number 52: 7 observations

predicted class=0 expected loss=0.1428571 P(node) =0.009803922

class counts: 6 1

probabilities: 0.857 0.143

Node number 53: 52 observations, complexity param=0.01034483

predicted class=1 expected loss=0.4807692 P(node) =0.07282913

class counts: 25 27

probabilities: 0.481 0.519

left son=106 (39 obs) right son=107 (13 obs)

Primary splits:

Parch < 0.5 to the left, improve=1.0384620, (0 missing)

SibSp < 0.5 to the right, improve=0.5833872, (0 missing)

Age < 21.5 to the left, improve=0.5833872, (0 missing)

Fare < 7.8875 to the right, improve=0.5170940, (0 missing)

Embarked splits as -LLR, improve=0.3520147, (0 missing)

Surrogate splits:

Fare < 18.62915 to the left, agree=0.808, adj=0.231, (0 split)

Embarked splits as -RLL, agree=0.788, adj=0.154, (0 split)

Node number 106: 39 observations, complexity param=0.01034483

predicted class=0 expected loss=0.4615385 P(node) =0.05462185

class counts: 21 18

probabilities: 0.538 0.462

left son=212 (23 obs) right son=213 (16 obs)

Primary splits:

Fare < 7.8875 to the right, improve=1.4498330, (0 missing)

Age < 18.5 to the left, improve=0.5274725, (0 missing)

SibSp < 0.5 to the right, improve=0.2937063, (0 missing)

Surrogate splits:

Embarked splits as -LRL, agree=0.718, adj=0.312, (0 split)

Age < 22.5 to the right, agree=0.641, adj=0.125, (0 split)

SibSp < 0.5 to the right, agree=0.641, adj=0.125, (0 split)

Node number 107: 13 observations

predicted class=1 expected loss=0.3076923 P(node) =0.01820728

class counts: 4 9

probabilities: 0.308 0.692

Node number 212: 23 observations

predicted class=0 expected loss=0.3478261 P(node) =0.03221289

class counts: 15 8

probabilities: 0.652 0.348

Node number 213: 16 observations

predicted class=1 expected loss=0.375 P(node) =0.02240896

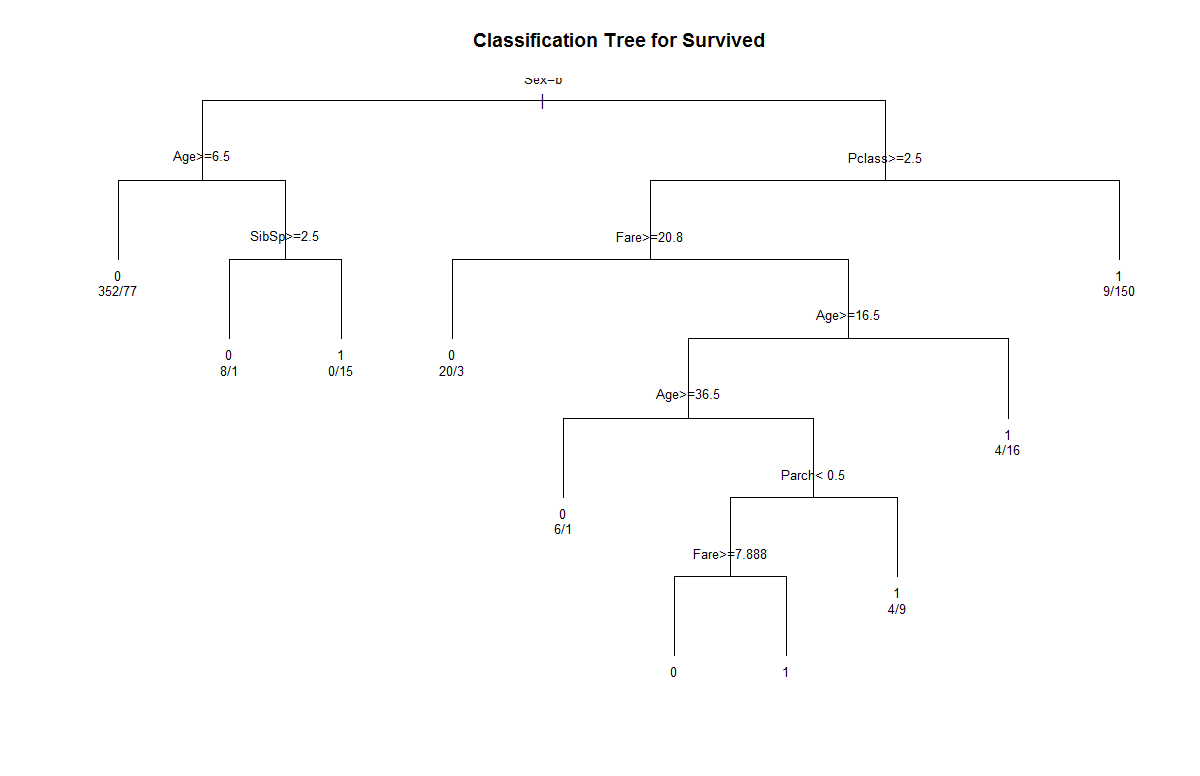
class counts: 6 10

probabilities: 0.375 0.625

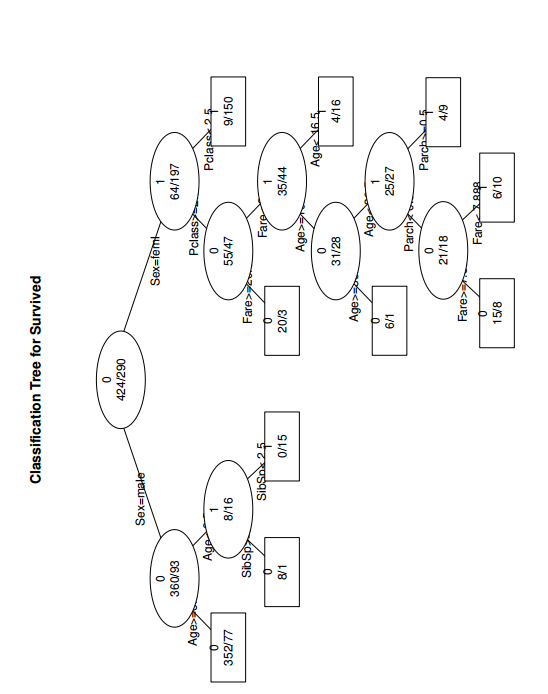
> #plot tree

> plot(fit, uniform = TRUE, main= "Classification Tree for Survived")

> text(fit, use.n = TRUE, cex =.8)



> post(fit, file= "C:/Users/Wanwan Zhang/Desktop/2016FALL/ADS/traintree.ps", title = "Classification Tree for Survived")



> # prune the tree

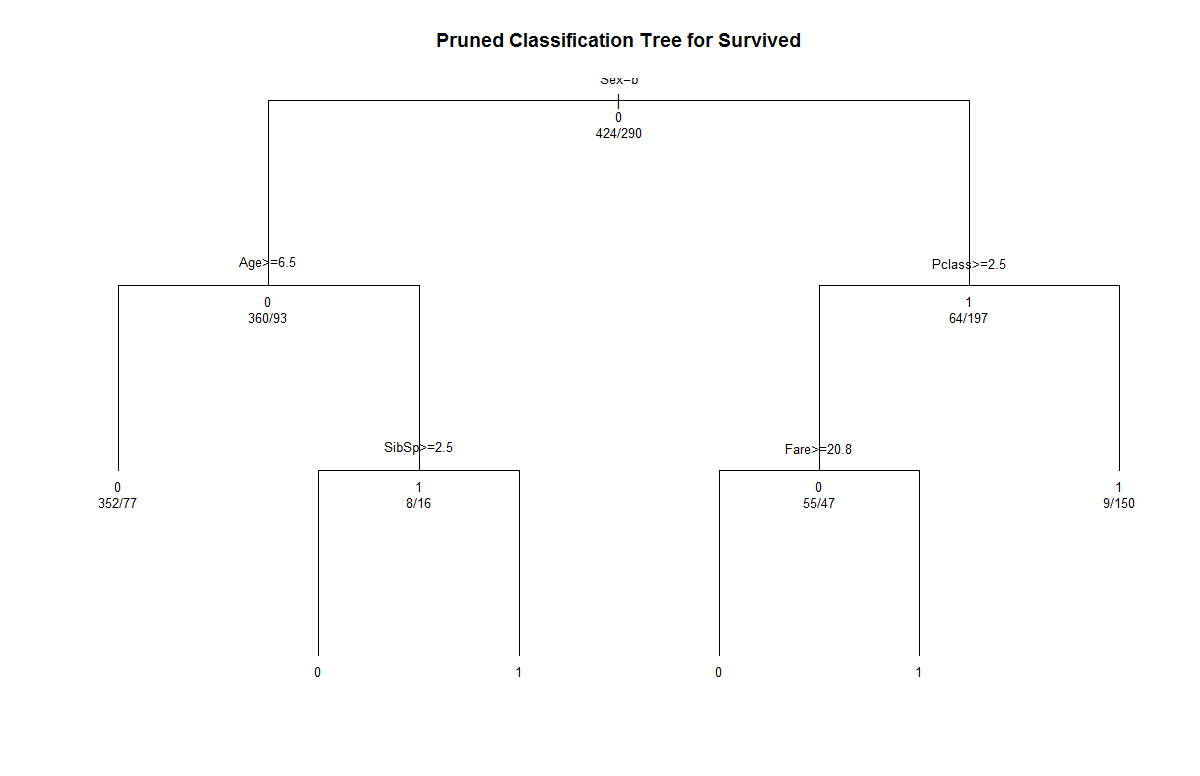
> pfit<- prune(fit, cp= fit$cptable[which.min(fit$cptable[,"xerror"]),"CP"])

> # plot the pruned tree

> plot(pfit, uniform=TRUE,

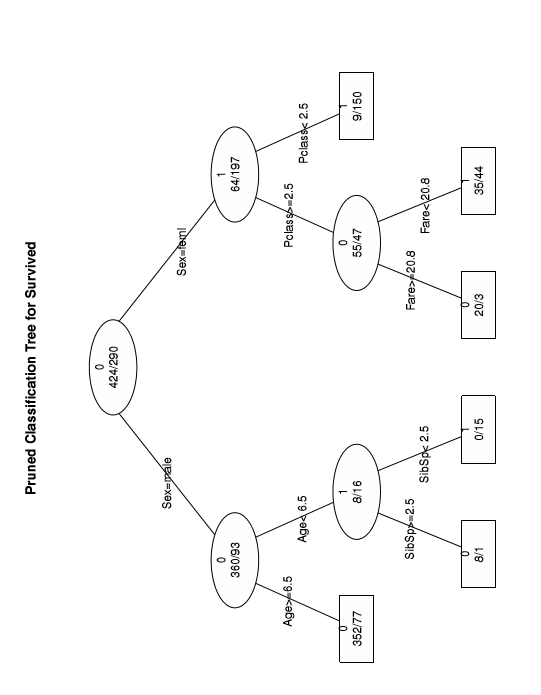
+ main="Pruned Classification Tree for Survived")

> text(pfit, use.n=TRUE, all=TRUE, cex=.8)



> post(pfit, file = "C:/Users/Wanwan Zhang/Desktop/2016FALL/ADS/ptraintree.ps",

+ title = "Pruned Classification Tree for Survived")



**Decision Tree Regression Energy efficiency**

> library(rpart)

> regressiontest<-read.csv("C:/Users/Wanwan Zhang/Desktop/2016FALL/ADS/5/ENB2012\_data.csv")

> # grow tree

> fit <- rpart(Y1~X1 + X2 + X3 + X4+X5 + X6 + X7 + X8,

+ method="anova", data=regressiontest)

> printcp(fit) # display the results

Regression tree:

rpart(formula = Y1 ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8, data = regressiontest,

method = "anova")

Variables actually used in tree construction:

[1] X1 X7

Root node error: 78090/768 = 101.68

n=768 (528 observations deleted due to missingness)

CP nsplit rel error xerror xstd

1 0.791087 0 1.000000 1.000440 0.0313809

2 0.084437 1 0.208913 0.209325 0.0125577

3 0.028959 2 0.124476 0.125436 0.0082240

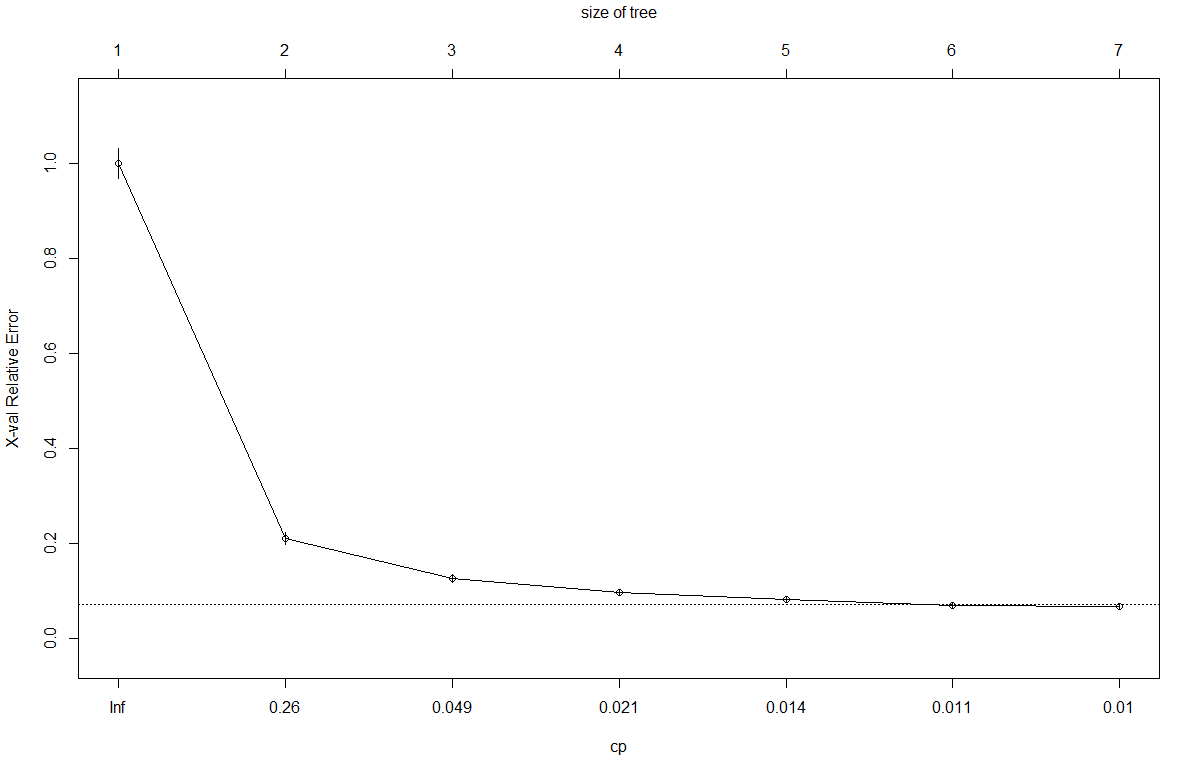
4 0.014749 3 0.095517 0.096395 0.0061341

5 0.012701 4 0.080768 0.081416 0.0046529

6 0.010142 5 0.068067 0.068930 0.0041404

7 0.010000 6 0.057925 0.066459 0.0039657

> plotcp(fit) # visualize cross-validation results



> summary(fit) # detailed summary of splits

Call:

rpart(formula = Y1 ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8, data = regressiontest,

method = "anova")

n=768 (528 observations deleted due to missingness)

CP nsplit rel error xerror xstd

1 0.79108692 0 1.00000000 1.00043994 0.031380899

2 0.08443700 1 0.20891308 0.20932546 0.012557688

3 0.02895925 2 0.12447608 0.12543591 0.008224005

4 0.01474895 3 0.09551682 0.09639492 0.006134131

5 0.01270065 4 0.08076787 0.08141554 0.004652873

6 0.01014180 5 0.06806722 0.06892997 0.004140406

7 0.01000000 6 0.05792542 0.06645926 0.003965688

Variable importance

X1 X2 X4 X5 X3 X7 X8

23 23 21 21 9 2 1

Node number 1: 768 observations, complexity param=0.7910869

mean=22.3072, MSE=101.6795

left son=2 (384 obs) right son=3 (384 obs)

Primary splits:

X1 < 0.75 to the left, improve=0.7910869, (0 missing)

X2 < 673.75 to the right, improve=0.7910869, (0 missing)

X4 < 183.75 to the right, improve=0.7910869, (0 missing)

X5 < 5.25 to the left, improve=0.7910869, (0 missing)

X3 < 281.75 to the left, improve=0.2104524, (0 missing)

Surrogate splits:

X2 < 673.75 to the right, agree=1.000, adj=1.000, (0 split)

X4 < 183.75 to the right, agree=1.000, adj=1.000, (0 split)

X5 < 5.25 to the left, agree=1.000, adj=1.000, (0 split)

X3 < 281.75 to the left, agree=0.667, adj=0.333, (0 split)

Node number 2: 384 observations, complexity param=0.01270065

mean=13.33852, MSE=7.119553

left son=4 (144 obs) right son=5 (240 obs)

Primary splits:

X7 < 0.175 to the left, improve=0.3627744, (0 missing)

X1 < 0.65 to the right, improve=0.3128987, (0 missing)

X2 < 771.75 to the left, improve=0.3128987, (0 missing)

X3 < 330.75 to the left, improve=0.3128987, (0 missing)

X8 < 0.5 to the left, improve=0.3093887, (0 missing)

Surrogate splits:

X8 < 0.5 to the left, agree=0.688, adj=0.167, (0 split)

Node number 3: 384 observations, complexity param=0.084437

mean=31.27589, MSE=35.36479

left son=6 (256 obs) right son=7 (128 obs)

Primary splits:

X1 < 0.805 to the right, improve=0.4855400, (0 missing)

X2 < 624.75 to the left, improve=0.4855400, (0 missing)

X3 < 330.75 to the left, improve=0.4855400, (0 missing)

X7 < 0.175 to the left, improve=0.2513315, (0 missing)

X8 < 0.5 to the left, improve=0.1997350, (0 missing)

Surrogate splits:

X2 < 624.75 to the left, agree=1, adj=1, (0 split)

X3 < 330.75 to the left, agree=1, adj=1, (0 split)

Node number 4: 144 observations

mean=11.26375, MSE=5.476593

Node number 5: 240 observations

mean=14.58338, MSE=3.972863

Node number 6: 256 observations, complexity param=0.02895925

mean=28.34578, MSE=19.12815

left son=12 (96 obs) right son=13 (160 obs)

Primary splits:

X7 < 0.175 to the left, improve=0.46181610, (0 missing)

X8 < 0.5 to the left, improve=0.36685200, (0 missing)

X1 < 0.84 to the left, improve=0.13569390, (0 missing)

X2 < 600.25 to the right, improve=0.13569390, (0 missing)

X4 < 134.75 to the right, improve=0.08723999, (0 missing)

Surrogate splits:

X8 < 0.5 to the left, agree=0.688, adj=0.167, (0 split)

Node number 7: 128 observations, complexity param=0.01474895

mean=37.13609, MSE=16.32503

left son=14 (48 obs) right son=15 (80 obs)

Primary splits:

X7 < 0.175 to the left, improve=0.5511780, (0 missing)

X8 < 0.5 to the left, improve=0.4383955, (0 missing)

X1 < 0.775 to the left, improve=0.1327052, (0 missing)

X2 < 649.25 to the right, improve=0.1327052, (0 missing)

X3 < 379.75 to the right, improve=0.1327052, (0 missing)

Surrogate splits:

X8 < 0.5 to the left, agree=0.688, adj=0.167, (0 split)

Node number 12: 96 observations, complexity param=0.0101418

mean=24.50875, MSE=12.73365

left son=24 (16 obs) right son=25 (80 obs)

Primary splits:

X7 < 0.05 to the left, improve=0.64786610, (0 missing)

X8 < 0.5 to the left, improve=0.64786610, (0 missing)

X1 < 0.84 to the left, improve=0.10668190, (0 missing)

X2 < 600.25 to the right, improve=0.10668190, (0 missing)

X4 < 116.375 to the left, improve=0.07098019, (0 missing)

Surrogate splits:

X8 < 0.5 to the left, agree=1, adj=1, (0 split)

Node number 13: 160 observations

mean=30.648, MSE=8.830945

Node number 14: 48 observations

mean=33.26354, MSE=12.453

Node number 15: 80 observations

mean=39.45963, MSE=4.251449

Node number 24: 16 observations

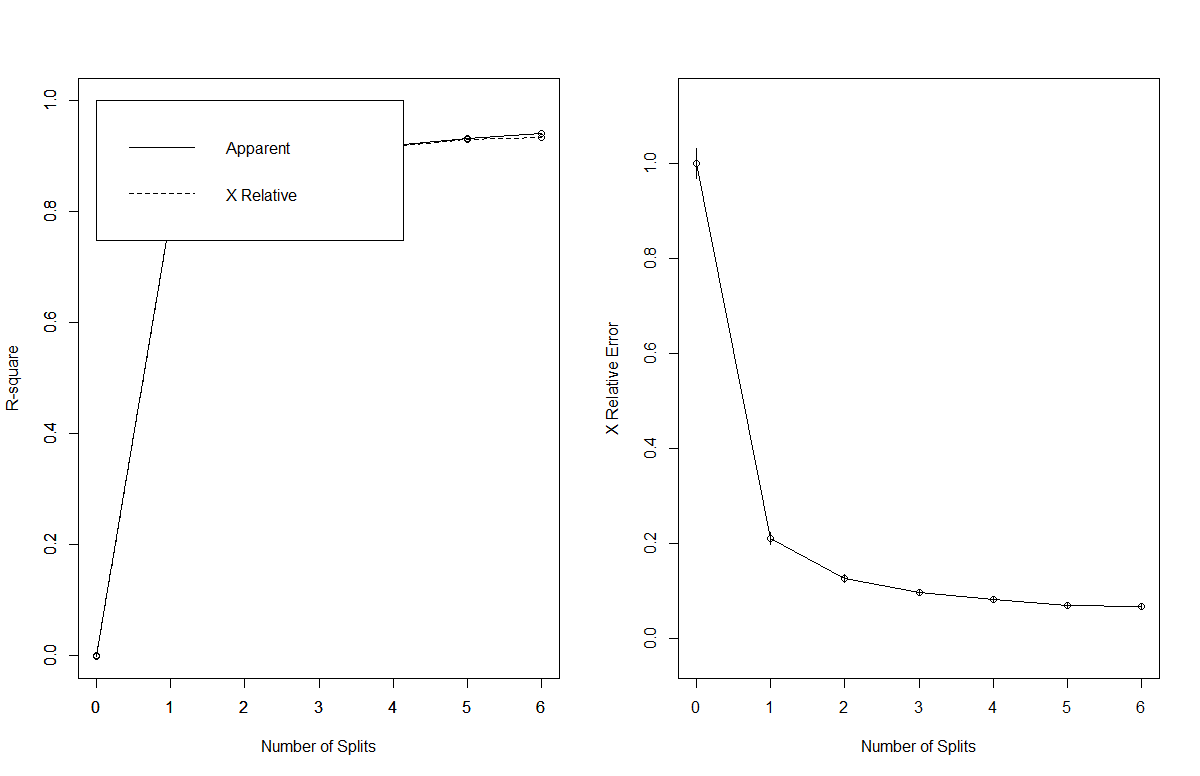
mean=18.08625, MSE=4.279523

Node number 25: 80 observations

mean=25.79325, MSE=4.524837

> # create additional plots

> par(mfrow=c(1,2)) # two plots on one page



> rsq.rpart(fit) # visualize cross-validation results

Regression tree:

rpart(formula = Y1 ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8, data = regressiontest,

method = "anova")

Variables actually used in tree construction:

[1] X1 X7

Root node error: 78090/768 = 101.68

n=768 (528 observations deleted due to missingness)

CP nsplit rel error xerror xstd

1 0.791087 0 1.000000 1.000440 0.0313809

2 0.084437 1 0.208913 0.209325 0.0125577

3 0.028959 2 0.124476 0.125436 0.0082240

4 0.014749 3 0.095517 0.096395 0.0061341

5 0.012701 4 0.080768 0.081416 0.0046529

6 0.010142 5 0.068067 0.068930 0.0041404

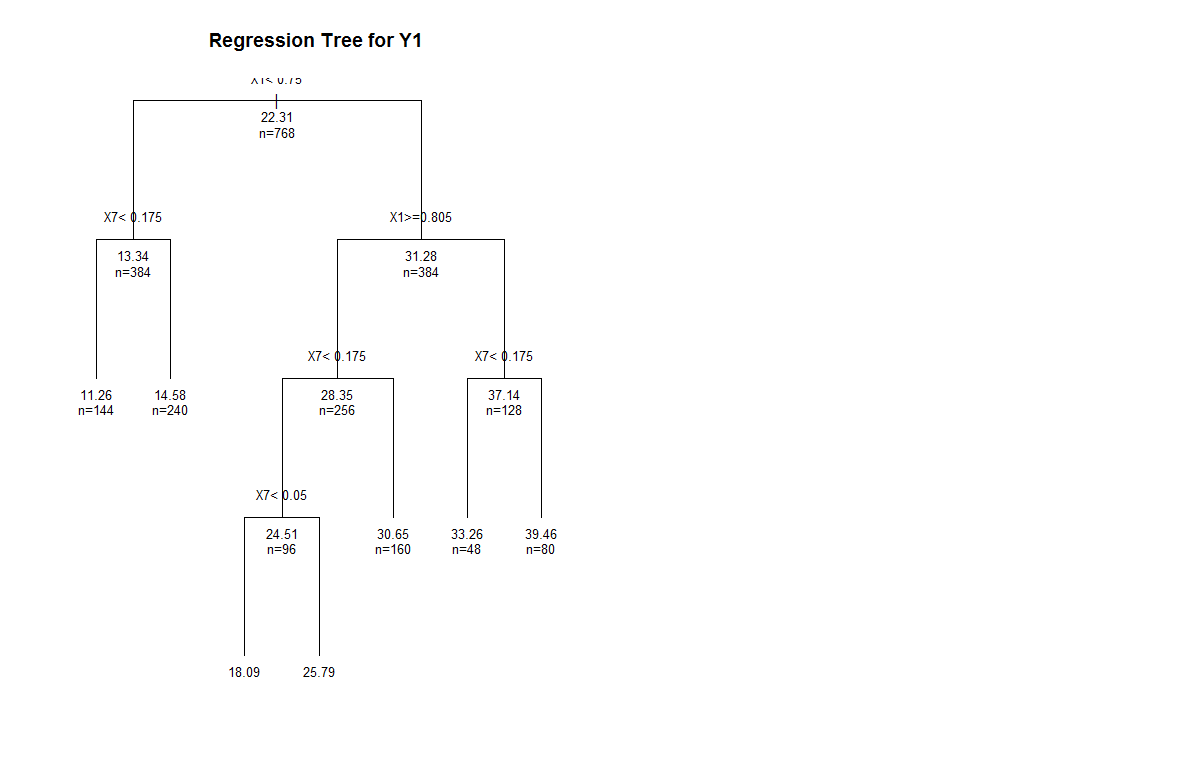
7 0.010000 6 0.057925 0.066459 0.0039657

> # plot tree

> plot(fit, uniform=TRUE,

+ main="Regression Tree for Y1 ")

> text(fit, use.n=TRUE, all=TRUE, cex=.8)



> # create attractive postcript plot of tree

> post(fit, file = "c:/tree2.ps",

+ title = "Regression Tree for Y1 ")

> # prune the tree

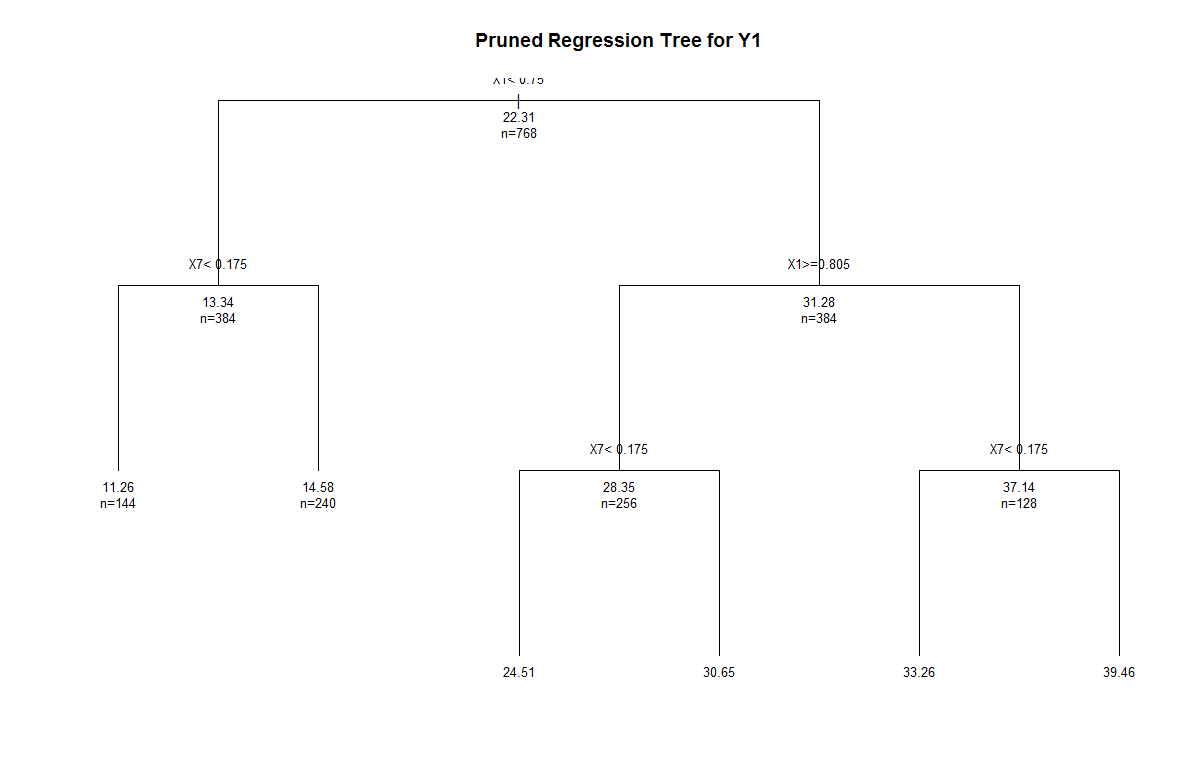
> pfit<- prune(fit, cp=fit$cptable[which.min(fit$cptable[,"xerror"]),"CP"]) # from cptable

> # plot the pruned tree

> plot(pfit, uniform=TRUE,

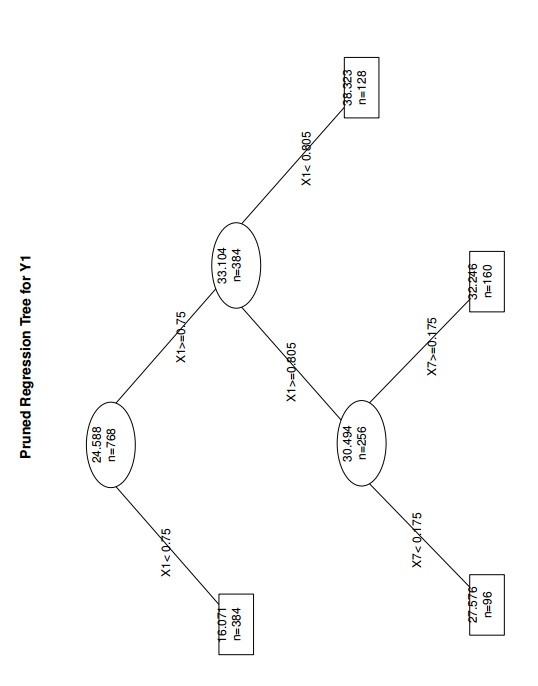
+ main="Pruned Regression Tree for Y1")

> text(pfit, use.n=TRUE, all=TRUE, cex=.8)



> post(pfit, file = "c:/ptree2.ps",

+ title = "Pruned Regression Tree for Y1")



> # grow tree

> fit <- rpart(Y2~X1 + X2 + X3 + X4+X5 + X6 + X7 + X8,

+ method="anova", data=regressiontest)

> printcp(fit) # display the results

Regression tree:

rpart(formula = Y2 ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8, data = regressiontest,

method = "anova")

Variables actually used in tree construction:

[1] X1 X7

Root node error: 69416/768 = 90.385

n=768 (528 observations deleted due to missingness)

CP nsplit rel error xerror xstd

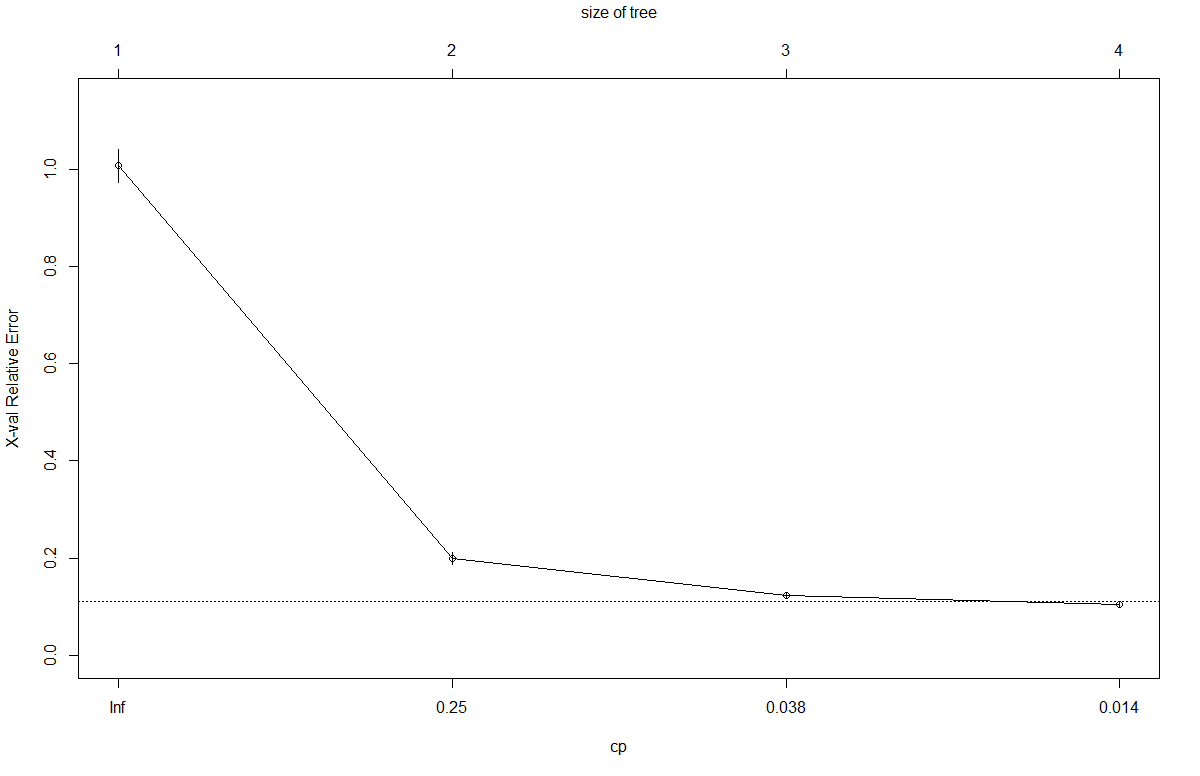
1 0.802431 0 1.00000 1.00711 0.0335077

2 0.075344 1 0.19757 0.19882 0.0126383

3 0.018851 2 0.12222 0.12331 0.0067774

4 0.010000 3 0.10337 0.10439 0.0056317

> plotcp(fit) # visualize cross-validation results



> summary(fit) # detailed summary of splits

Call:

rpart(formula = Y2 ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8, data = regressiontest,

method = "anova")

n=768 (528 observations deleted due to missingness)

CP nsplit rel error xerror xstd

1 0.80243107 0 1.0000000 1.0071084 0.033507743

2 0.07534417 1 0.1975689 0.1988189 0.012638269

3 0.01885084 2 0.1222248 0.1233132 0.006777369

4 0.01000000 3 0.1033739 0.1043860 0.005631741

Variable importance

X1 X2 X4 X5 X3 X7

24 24 22 22 9 1

Node number 1: 768 observations, complexity param=0.8024311

mean=24.58776, MSE=90.38514

left son=2 (384 obs) right son=3 (384 obs)

Primary splits:

X1 < 0.75 to the left, improve=0.8024311, (0 missing)

X2 < 673.75 to the right, improve=0.8024311, (0 missing)

X4 < 183.75 to the right, improve=0.8024311, (0 missing)

X5 < 5.25 to the left, improve=0.8024311, (0 missing)

X3 < 281.75 to the left, improve=0.2067026, (0 missing)

Surrogate splits:

X2 < 673.75 to the right, agree=1.000, adj=1.000, (0 split)

X4 < 183.75 to the right, agree=1.000, adj=1.000, (0 split)

X5 < 5.25 to the left, agree=1.000, adj=1.000, (0 split)

X3 < 281.75 to the left, agree=0.667, adj=0.333, (0 split)

Node number 2: 384 observations

mean=16.07143, MSE=5.844996

Node number 3: 384 observations, complexity param=0.07534417

mean=33.10409, MSE=29.8696

left son=6 (256 obs) right son=7 (128 obs)

Primary splits:

X1 < 0.805 to the right, improve=0.4559816, (0 missing)

X2 < 624.75 to the left, improve=0.4559816, (0 missing)

X3 < 330.75 to the left, improve=0.4559816, (0 missing)

X7 < 0.175 to the left, improve=0.1645447, (0 missing)

X4 < 116.375 to the left, improve=0.1011679, (0 missing)

Surrogate splits:

X2 < 624.75 to the left, agree=1, adj=1, (0 split)

X3 < 330.75 to the left, agree=1, adj=1, (0 split)

Node number 6: 256 observations, complexity param=0.01885084

mean=30.49449, MSE=15.61716

left son=12 (96 obs) right son=13 (160 obs)

Primary splits:

X7 < 0.175 to the left, improve=0.32730070, (0 missing)

X8 < 0.5 to the left, improve=0.17424270, (0 missing)

X1 < 0.84 to the left, improve=0.12983510, (0 missing)

X2 < 600.25 to the right, improve=0.12983510, (0 missing)

X4 < 134.75 to the right, improve=0.06719687, (0 missing)

Surrogate splits:

X8 < 0.5 to the left, agree=0.688, adj=0.167, (0 split)

Node number 7: 128 observations

mean=38.32328, MSE=17.51451

Node number 12: 96 observations

mean=27.57573, MSE=10.59244

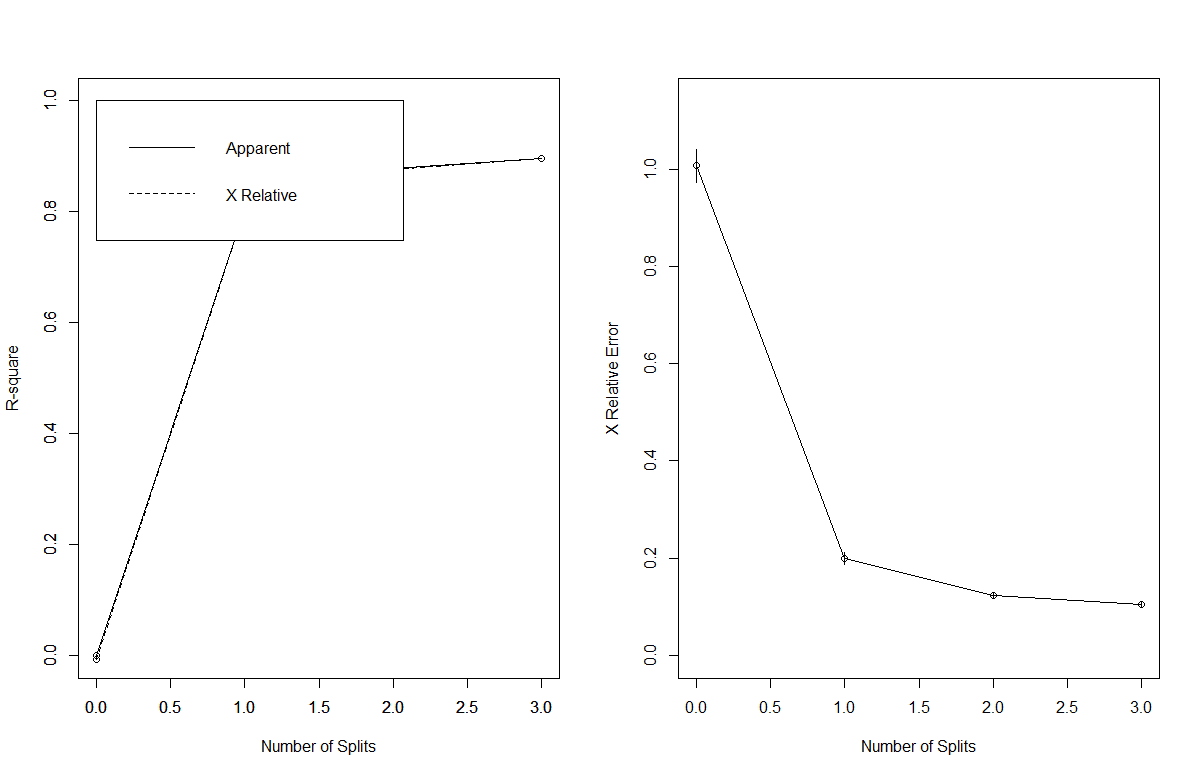
Node number 13: 160 observations

mean=32.24575, MSE=10.45358

> # create additional plots

> par(mfrow=c(1,2)) # two plots on one page

> rsq.rpart(fit) # visualize cross-validation results



Regression tree:

rpart(formula = Y2 ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8, data = regressiontest,

method = "anova")

Variables actually used in tree construction:

[1] X1 X7

Root node error: 69416/768 = 90.385

n=768 (528 observations deleted due to missingness)

CP nsplit rel error xerror xstd

1 0.802431 0 1.00000 1.00711 0.0335077

2 0.075344 1 0.19757 0.19882 0.0126383

3 0.018851 2 0.12222 0.12331 0.0067774

4 0.010000 3 0.10337 0.10439 0.0056317

> # plot tree

> plot(fit, uniform=TRUE,

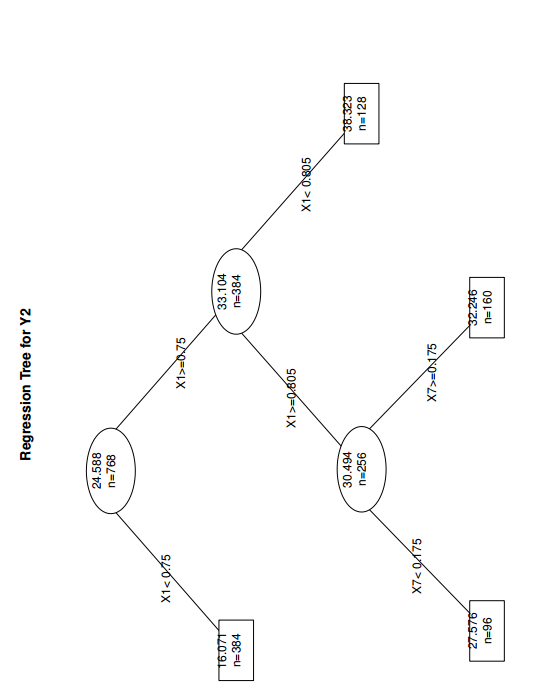
+ main="Regression Tree for Y2 ")

> text(fit, use.n=TRUE, all=TRUE, cex=.8)

> # create attractive postcript plot of tree

> post(fit, file = "c:/tree3.ps",

+ title = "Regression Tree for Y2 ")



> # prune the tree

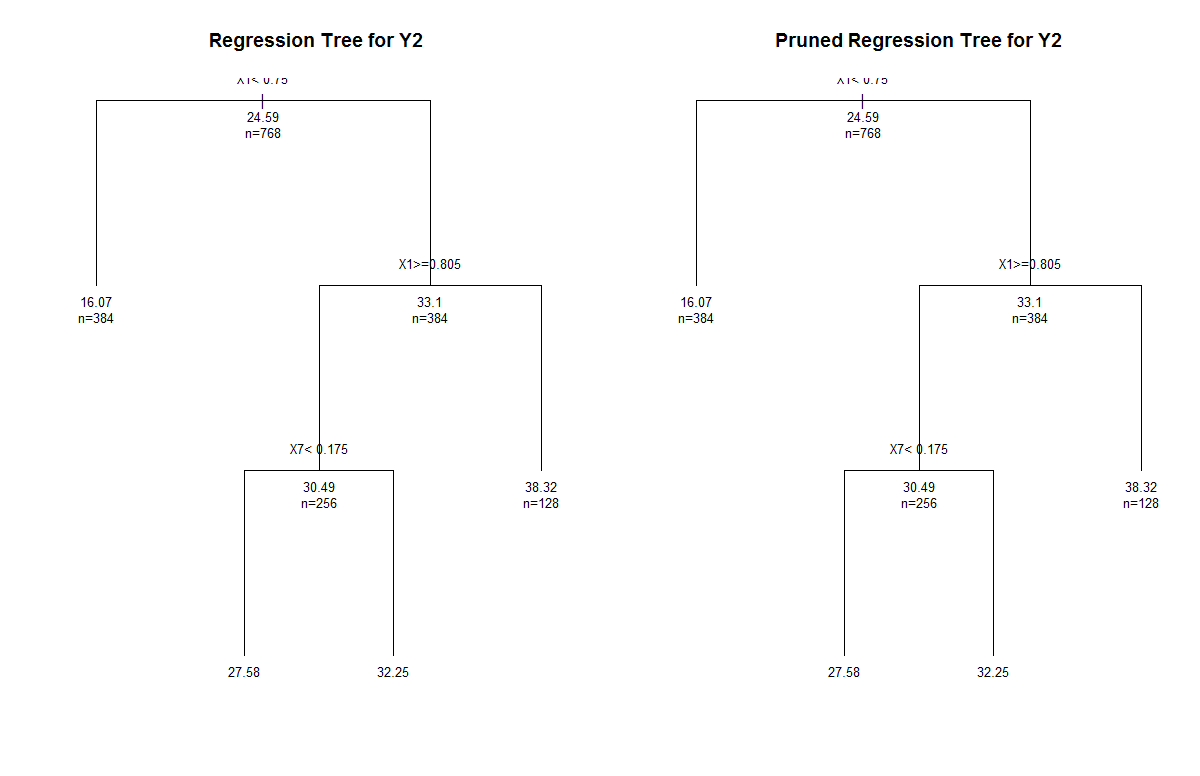
> pfit<- prune(fit, cp=fit$cptable[which.min(fit$cptable[,"xerror"]),"CP"]) # from cptable

> # plot the pruned tree

> plot(pfit, uniform=TRUE,

+ main="Pruned Regression Tree for Y2")

> text(pfit, use.n=TRUE, all=TRUE, cex=.8)



> post(pfit, file = " C:/Users/Wanwan Zhang/Desktop/2016FALL/ADS/5/PTREE3.PS ",

+ title = "Pruned Regression Tree for Y2")

