14.1

Problem Statement

1. Use the below given data set

DataSet

Problem- prediction of the number of comments in the upcoming 24 hours on

those blogs, The train data was generated from different base times that may

temporally overlap. Therefore, if you simply split the train into disjoint partitions,

the underlying time intervals may overlap. Therefore, the you should use the

provided, temporally disjoint train and test splits to ensure that the evaluation is

fair.

a. Read the dataset and identify the right features

b. Clean dataset, impute missing values and perform exploratory data analysis.

c. Visualize the dataset and make inferences from that

d. Perform any 3 hypothesis tests using columns of your choice, make conclusions

e. Create a linear regression model to predict the no.of comments in the next 24hrs (relative to basetime)

14.2

Problem- prediction of the number of comments in the upcoming 24 hours on

those blogs, The train data was generated from different base times that may

temporally overlap. Therefore, if you simply split the train into disjoint partitions,

the underlying time intervals may overlap. Therefore, the you should use the

provided, temporally disjoint train and test splits to ensure that the evaluation is

fair.

a. Read the dataset and identify the right features

b. Clean dataset, impute missing values and perform exploratory data analysis.

c. Visualize the dataset and make inferences from that

d. Perform any 3 hypothesis tests using columns of your choice, make conclusions

e. Create a linear regression model to predict the number of comments in the next 24 hours

(relative to basetime)

> summary(mod)

Call:

lm(formula = blogData\_train$A ~ blogData\_train$Target)

Residuals:

Min 1Q Median 3Q Max

-1311.76 -30.97 -21.92 6.08 1089.09

Coefficients:

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

(Intercept) 32.553104 0.307019 106.0 <2e-16 \*\*\*

blogData\_train$Target 1.018677 0.008014 127.1 <2e-16 \*\*\*

---

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

Residual standard error: 69.17 on 52395 degrees of freedom

Multiple R-squared: 0.2357, Adjusted R-squared: 0.2357

F-statistic: 1.616e+04 on 1 and 52395 DF, p-value: < 2.2e-16

**For 1 unit increase there will be 1.01 unit of increase in no.of comments**

**These 3 terms, t-statistics (127.1), Std Error of regression coefficient (0.008014) and Probability (of committing Type I Error) is 2e-06 [2/1000000 = 0.00000200 = almost 0] are associated with a t test which tests following Null Hypothesis:**

**Ho: The slope of no.of comments is not significant**

**Ha: The slope of no.of comments is significant**

**0.008014 = standard error of regression co-efficient**

**As p value is 0.000 which is less than 0.05, Reject the Ho (and accept Ha) and conclude that “slope is significant”**

**Multiple R-squared = 0.2357**

**23.5% of the variance in Average can be explained by no.of comments [Remaining 76.5% is unexplained variance....due to factors outside the model]**

> pred<- predict(lm(blogData\_train$A~blogData\_train$Target))

> pred

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15

33.57178 32.55310 32.55310 33.57178 60.05738 32.55310 32.55310 60.05738 41.72120 41.72120 32.55310 32.55310 34.59046 32.55310 45.79590

16 17 18 19 20 21 22 23 24 25 26 27 28 29 30

47.83326 35.60913 47.83326 45.79590 33.57178 53.94532 33.57178 53.94532 33.57178 35.60913 35.60913 35.60913 35.60913 34.59046 34.59046

31 32 33 34 35 36 37 38 39 40 41 42 43 44 45

37.64649 37.64649 32.55310 32.55310 53.94532 32.55310 32.55310 35.60913 32.55310 32.55310 32.55310 34.59046 32.55310 32.55310 32.55310

46 47 48 49 50 51 52 53 54 55 56 57 58 59 60

32.55310 46.81458 44.77723 39.68384 32.55310 32.55310 32.55310 32.55310 32.55310 32.55310 32.55310 32.55310 32.55310 39.68384 35.60913

61 62 63 64 65 66 67 68 69 70 71 72 73 74 75

33.57178 33.57178 33.57178 34.59046 34.59046 32.55310 32.55310 37.64649 49.87061 37.64649 32.55310 34.59046 32.55310 33.57178 32.55310

76 77 78 79 80 81 82 83 84 85 86 87 88 89 90

33.57178 40.70252 32.55310 32.55310 69.22547 33.57178 34.59046 44.77723 41.72120 40.70252 41.72120 40.70252 36.62781 40.70252 33.57178

91 92 93 94 95 96 97 98 99 100 101 102 103 104 105

33.57178 35.60913 35.60913 34.59046 34.59046 33.57178 33.57178 35.60913 35.60913 34.59046 34.59046 35.60913 35.60913 59.03870 34.59046

106 107 108 109 110 111 112 113 114 115 116 117 118 119 120

33.57178 37.64649 33.57178 35.60913 34.59046 35.60913 57.00135 57.00135 34.59046 34.59046 35.60913 33.57178 104.87916 32.55310 32.55310

121 122 123 124 125 126 127 128 129 130 131 132 133 134 135

36.62781 37.64649 37.64649 37.64649 36.62781 36.62781 36.62781 38.66517 38.66517 38.66517 78.39356 78.39356 40.70252 40.70252 32.55310

136 137 138 139 140 141 142 143 144 145 146 147 148 149 150

32.55310 36.62781 36.62781 39.68384 32.55310 32.55310 32.55310 32.55310 32.55310 32.55310 44.77723 32.55310 38.66517 33.57178 36.62781

151 152 153 154 155 156 157 158 159 160 161 162 163 164 165

33.57178 32.55310 32.55310 38.66517 32.55310 38.66517 34.59046 34.59046 39.68384 39.68384 37.64649 37.64649 36.62781 36.62781 33.57178

166 167 168 169 170 171 172 173 174 175 176 177 178 179 180

33.57178 33.57178 32.55310 107.93519 32.55310 107.93519 52.92664 52.92664 37.64649 37.64649 50.88929 50.88929 33.57178 33.57178 40.70252

181 182 183 184 185 186 187 188 189 190 191 192 193 194 195

32.55310 32.55310 34.59046 32.55310 32.55310 38.66517 32.55310 32.55310 32.55310 34.59046 33.57178 33.57178 32.55310 32.55310 32.55310

196 197 198 199 200 201 202 203 204 205 206 207 208 209 210

33.57178 33.57178 32.55310 32.55310 32.55310 32.55310 32.55310 32.55310 32.55310 32.55310 32.55310 32.55310 55.98267 40.70252 37.64649

211 212 213 214 215 216 217 218 219 220 221 222 223 224 225

36.62781 36.62781 36.62781 36.62781 32.55310 35.60913 32.55310 35.60913 36.62781 36.62781 80.43091 80.43091 33.57178 80.43091 33.57178

226 227 228 229 230 231 232 233 234 235 236 237 238 239 240

36.62781 33.57178 33.57178 36.62781 112.00989 112.00989 112.00989 33.57178 32.55310 32.55310 32.55310 32.55310 41.72120 41.72120 41.72120

241 242 243 244 245 246 247 248 249 250 251 252 253 254 255

32.55310 32.55310 32.55310 35.60913 35.60913 35.60913 33.57178 33.57178 33.57178 33.57178 35.60913 32.55310 32.55310 35.60913 35.60913

256 257 258 259 260 261 262 263 264 265 266 267 268 269 270

33.57178 33.57178 33.57178 32.55310 32.55310 32.55310 32.55310 32.55310 32.55310 39.68384 39.68384 39.68384 36.62781 36.62781 36.62781

271 272 273 274 275 276 277 278 279 280 281 282 283 284 285

39.68384 39.68384 39.68384 35.60913 35.60913 33.57178 33.57178 44.77723 32.55310 32.55310 44.77723 38.66517 38.66517 36.62781 36.62781

286 287 288 289 290 291 292 293 294 295 296 297 298 299 300

38.66517 38.66517 38.66517 33.57178 33.57178 33.57178 35.60913 35.60913 35.60913 36.62781 36.62781 35.60913 35.60913 32.55310 32.55310

301 302 303 304 305 306 307 308 309 310 311 312 313 314 315

34.59046 34.59046 34.59046 43.75855 43.75855 35.60913 35.60913 35.60913 34.59046 34.59046 32.55310 32.55310 32.55310 34.59046 34.59046

316 317 318 319 320 321 322 323 324 325 326 327 328 329 330

35.60913 35.60913 33.57178 33.57178 32.55310 32.55310 41.72120 41.72120 41.72120 39.68384 39.68384 41.72120 34.59046 69.22547 69.22547

331 332 333 334 335 336 337 338 339 340 341 342 343 344 345

34.59046 34.59046 34.59046 42.73987 40.70252 34.59046 40.70252 32.55310 34.59046 32.55310 34.59046 34.59046 41.72120 41.72120 40.70252

346 347 348 349 350 351 352 353 354 355 356 357 358 359 360

39.68384 39.68384 40.70252 32.55310 37.64649 37.64649 32.55310 32.55310 33.57178 33.57178 32.55310 32.55310 33.57178 32.55310 35.60913

361 362 363 364 365 366 367 368 369 370 371 372 373 374 375

32.55310 35.60913 33.57178 88.58033 88.58033 46.81458 46.81458 32.55310 32.55310 40.70252 36.62781 33.57178 43.75855 34.59046 32.55310

376 377 378 379 380 381 382 383 384 385 386 387 388 389 390

36.62781 34.59046 33.57178 34.59046 34.59046 32.55310 32.55310 32.55310 32.55310 42.73987 42.73987 42.73987 42.73987 51.90796 37.64649

391 392 393 394 395 396 397 398 399 400 401 402 403 404 405

51.90796 47.83326 37.64649 47.83326 47.83326 35.60913 36.62781 36.62781 38.66517 35.60913 35.60913 35.60913 35.60913 38.66517 35.60913

406 407 408 409 410 411 412 413 414 415 416 417 418 419 420

33.57178 40.70252 40.70252 42.73987 33.57178 32.55310 32.55310 32.55310 42.73987 33.57178 32.55310 36.62781 32.55310 36.62781 32.55310

421 422 423 424 425 426 427 428 429 430 431 432 433 434 435

37.64649 35.60913 42.73987 32.55310 32.55310 32.55310 32.55310 32.55310 32.55310 34.59046 34.59046 40.70252 51.90796 40.70252 51.90796

436 437 438 439 440 441 442 443 444 445 446 447 448 449 450

33.57178 36.62781 33.57178 36.62781 41.72120 41.72120 32.55310 32.55310 38.66517 38.66517 33.57178 33.57178 33.57178 33.57178 34.59046

451 452 453 454 455 456 457 458 459 460 461 462 463 464 465

34.59046 35.60913 35.60913 92.65503 92.65503 32.55310 32.55310 50.88929 50.88929 36.62781 32.55310 32.55310 36.62781 39.68384 38.66517

466 467 468 469 470 471 472 473 474 475 476 477 478 479 480

38.66517 36.62781 39.68384 32.55310 32.55310 36.62781 37.64649 35.60913 37.64649 35.60913 38.66517 38.66517 158.86902 158.86902 158.86902

481 482 483 484 485 486 487 488 489 490 491 492 493 494 495

158.86902 88.58033 88.58033 88.58033 88.58033 53.94532 53.94532 53.94532 53.94532 41.72120 41.72120 33.57178 33.57178 33.57178 33.57178

496 497 498 499 500 501 502 503 504 505 506 507 508 509 510

64.13208 64.13208 39.68384 39.68384 35.60913 35.60913 34.59046 34.59046 32.55310 32.55310 40.70252 40.70252 36.62781 36.62781 36.62781

511 512 513 514 515 516 517 518 519 520 521 522 523 524 525

41.72120 41.72120 41.72120 35.60913 35.60913 35.60913 32.55310 32.55310 32.55310 32.55310 40.70252 40.70252 32.55310 40.70252 32.55310

526 527 528 529 530 531 532 533 534 535 536 537 538 539 540

32.55310 32.55310 32.55310 32.55310 32.55310 32.55310 41.72120 64.13208 42.73987 45.79590 45.79590 45.79590 44.77723 44.77723 44.77723

541 542 543 544 545 546 547 548 549 550 551 552 553 554 555

37.64649 37.64649 37.64649 35.60913 35.60913 72.28150 35.60913 35.60913 72.28150 72.28150 73.30018 38.66517 38.66517 73.30018 73.30018

556 557 558 559 560 561 562 563 564 565 566 567 568 569 570

48.85193 48.85193 48.85193 50.88929 37.64649 33.57178 32.55310 37.64649 34.59046 40.70252 47.83326 41.72120 33.57178 33.57178 33.57178

571 572 573 574 575 576 577 578 579 580 581 582 583 584 585

33.57178 34.59046 34.59046 42.73987 42.73987 42.73987 38.66517 38.66517 38.66517 33.57178 33.57178 33.57178 54.96399 54.96399 33.57178

586 587 588 589 590 591 592 593 594 595 596 597 598 599 600

33.57178 34.59046 34.59046 33.57178 33.57178 38.66517 32.55310 32.55310 38.66517 38.66517 38.66517 34.59046 32.55310 32.55310 34.59046

601 602 603 604 605 606 607 608 609 610 611 612 613 614 615

34.59046 34.59046 34.59046 34.59046 34.59046 34.59046 32.55310 32.55310 46.81458 33.57178 46.81458 33.57178 37.64649 32.55310 37.64649

616 617 618 619 620 621 622 623 624 625 626 627 628 629 630

32.55310 34.59046 34.59046 50.88929 50.88929 38.66517 39.68384 45.79590 45.79590 45.79590 39.68384 38.66517 33.57178 35.60913 32.55310

631 632 633 634 635 636 637 638 639 640 641 642 643 644 645

32.55310 32.55310 35.60913 33.57178 32.55310 32.55310 32.55310 32.55310 32.55310 39.68384 39.68384 33.57178 33.57178 33.57178 33.57178

646 647 648 649 650 651 652 653 654 655 656 657 658 659 660

33.57178 33.57178 32.55310 32.55310 32.55310 32.55310 71.26282 35.60913 36.62781 32.55310 53.94532 32.55310 43.75855 57.00135 42.73987

661 662 663 664 665 666 667 668 669 670 671 672 673 674 675

33.57178 33.57178 32.55310 32.55310 32.55310 32.55310 54.96399 54.96399 32.55310 32.55310 35.60913 35.60913 58.02002 74.31885 46.81458

676 677 678 679 680 681 682 683 684 685 686 687 688 689 690

34.59046 34.59046 32.55310 32.55310 33.57178 32.55310 33.57178 32.55310 32.55310 32.55310 32.55310 32.55310 34.59046 34.59046 34.59046

691 692 693 694 695 696 697 698 699 700 701 702 703 704 705

32.55310 32.55310 32.55310 34.59046 34.59046 34.59046 32.55310 34.59046 34.59046 32.55310 34.59046 65.15076 65.15076 32.55310 65.15076

706 707 708 709 710 711 712 713 714 715 716 717 718 719 720

42.73987 42.73987 42.73987 34.59046 34.59046 34.59046 38.66517 38.66517 38.66517 39.68384 34.59046 34.59046 39.68384 39.68384 32.55310

721 722 723 724 725 726 727 728 729 730 731 732 733 734 735

32.55310 32.55310 32.55310 33.57178 33.57178 33.57178 34.59046 34.59046 32.55310 34.59046 32.55310 32.55310 33.57178 33.57178 32.55310

736 737 738 739 740 741 742 743 744 745 746 747 748 749 750

33.57178 32.55310 32.55310 34.59046 34.59046 34.59046 36.62781 36.62781 36.62781 36.62781 36.62781 36.62781 32.55310 32.55310 32.55310

751 752 753 754 755 756 757 758 759 760 761 762 763 764 765

33.57178 33.57178 33.57178 34.59046 34.59046 34.59046 32.55310 32.55310 32.55310 34.59046 34.59046 34.59046 39.68384 39.68384 39.68384

766 767 768 769 770 771 772 773 774 775 776 777 778 779 780

32.55310 32.55310 32.55310 38.66517 38.66517 38.66517 32.55310 32.55310 32.55310 32.55310 32.55310 36.62781 36.62781 36.62781 32.55310

781 782 783 784 785 786 787 788 789 790 791 792 793 794 795

32.55310 44.77723 44.77723 44.77723 32.55310 32.55310 33.57178 57.00135 57.00135 57.00135 33.57178 33.57178 32.55310 32.55310 32.55310

796 797 798 799 800 801 802 803 804 805 806 807 808 809 810

38.66517 38.66517 38.66517 35.60913 35.60913 35.60913 32.55310 32.55310 32.55310 32.55310 32.55310 32.55310 35.60913 35.60913 32.55310

811 812 813 814 815 816 817 818 819 820 821 822 823 824 825

35.60913 32.55310 32.55310 33.57178 33.57178 33.57178 33.57178 33.57178 33.57178 33.57178 33.57178 33.57178 33.57178 33.57178 33.57178

826 827 828 829 830 831 832 833 834 835 836 837 838 839 840

34.59046 34.59046 34.59046 32.55310 32.55310 32.55310 32.55310 32.55310 32.55310 32.55310 32.55310 32.55310 32.55310 32.55310 32.55310

841 842 843 844 845 846 847 848 849 850 851 852 853 854 855

32.55310 32.55310 32.55310 33.57178 33.57178 32.55310 32.55310 32.55310 32.55310 32.55310 32.55310 33.57178 42.73987 42.73987 42.73987

856 857 858 859 860 861 862 863 864 865 866 867 868 869 870

37.64649 33.57178 33.57178 37.64649 33.57178 37.64649 35.60913 35.60913 35.60913 33.57178 33.57178 33.57178 32.55310 32.55310 32.55310

871 872 873 874 875 876 877 878 879 880 881 882 883 884 885

32.55310 32.55310 32.55310 33.57178 33.57178 33.57178 32.55310 32.55310 39.68384 32.55310 39.68384 32.55310 32.55310 35.60913 32.55310

886 887 888 889 890 891 892 893 894 895 896 897 898 899 900

35.60913 32.55310 32.55310 33.57178 32.55310 33.57178 33.57178 33.57178 35.60913 33.57178 35.60913 32.55310 32.55310 32.55310 32.55310

901 902 903 904 905 906 907 908 909 910 911 912 913 914 915

32.55310 32.55310 34.59046 32.55310 32.55310 34.59046 32.55310 34.59046 32.55310 33.57178 34.59046 34.59046 33.57178 33.57178 34.59046

916 917 918 919 920 921 922 923 924 925 926 927 928 929 930

32.55310 33.57178 33.57178 32.55310 32.55310 33.57178 34.59046 33.57178 33.57178 34.59046 33.57178 34.59046 33.57178 33.57178 33.57178

931 932 933 934 935 936 937 938 939 940 941 942 943 944 945

32.55310 44.77723 44.77723 44.77723 32.55310 32.55310 32.55310 32.55310 32.55310 36.62781 32.55310 32.55310 36.62781 32.55310 36.62781

946 947 948 949 950 951 952 953 954 955 956 957 958 959 960

32.55310 32.55310 32.55310 35.60913 35.60913 35.60913 33.57178 33.57178 33.57178 33.57178 55.98267 55.98267 33.57178 55.98267 33.57178

961 962 963 964 965 966 967 968 969 970 971 972 973 974 975

55.98267 34.59046 46.81458 46.81458 34.59046 46.81458 34.59046 46.81458 37.64649 37.64649 32.55310 37.64649 32.55310 32.55310 37.64649

976 977 978 979 980 981 982 983 984 985 986 987 988 989 990

32.55310 32.55310 32.55310 32.55310 32.55310 32.55310 32.55310 32.55310 32.55310 34.59046 32.55310 32.55310 34.59046 32.55310 32.55310

991 992 993 994 995 996 997 998 999 1000

34.59046 32.55310 32.55310 32.55310 32.55310 32.55310 32.55310 32.55310 32.55310 32.55310

[ reached getOption("max.print") -- omitted 51397 entries ]

f. Fine tune the model and represent important features

g. Interpret the summary of the linear model

h. Report the test accuracy vs. the training accuracy

i. Interpret the final model coefficients

j. Plot the model result and compare it with assumptions of the model

15.Problem Statement

1. Use the below given data set

DataSet

a. Predict the no of comments in next H hrs

b. Use regression technique

c. Report the training accuracy and test accuracy