TUGAS 3 DATA MINING

1. Membandingkan kebaikan model regresi yang dihasilkan
2. Regresi dengan data yang telah diatasi *Missing Value*nya

**Regression Analysis: Gross versus Ratings, Budget, ...**

The regression equation is

Gross = - 1.27E+08 + 17855771 Ratings + 0.840 Budget + 14247 Screens

- 1.44 Views + 568 Likes + 10468 Dislikes - 4298 Comments

+ 2.84 Aggregate Followers

Predictor Coef SE Coef T P

Constant -127015312 28889423 -4.40 0.000

Ratings 17855771 4431902 4.03 0.000

Budget 0.84045 0.09452 8.89 0.000

Screens 14247 3488 4.08 0.000

Views -1.437 1.677 -0.86 0.393

Likes 568.3 366.2 1.55 0.122

Dislikes 10468 5473 1.91 0.057

Comments -4298 3065 -1.40 0.162

Aggregate Followers 2.8387 0.8717 3.26 0.001

S = 56806979 R-Sq = 61.9% R-Sq(adj) = 60.4%

1. Regresi dengan data yang telah diatasi *Missing Value* dan *outlier*nya

**Regression Analysis: Gross versus Ratings, Budget, ...**

The regression equation is

Gross = - 1.10E+08 + 17055269 Ratings + 0.698 Budget + 12133 Screens

- 5.03 Views + 1263 Likes + 28696 Dislikes - 9981 Comments

+ 1.68 Aggregate Followers

Predictor Coef SE Coef T P

Constant -110305465 22745829 -4.85 0.000

Ratings 17055269 3510541 4.86 0.000

Budget 0.69767 0.07468 9.34 0.000

Screens 12133 2557 4.75 0.000

Views -5.026 1.874 -2.68 0.008

Likes 1263.0 592.8 2.13 0.034

Dislikes 28696 10424 2.75 0.006

Comments -9981 3798 -2.63 0.009

Aggregate Followers 1.682 1.024 1.64 0.102

S = 39980166 R-Sq = 65.3% R-Sq(adj) = 63.8%

1. Regresi dengan data yang telah diatasi *Missing Value*nya dan ditransformasikan

**Regression Analysis: Gross versus Ratings, Budget, ...**

The regression equation is

Gross = 49.6 - 11356 Ratings - 0.942 Budget + 13.1 Screens - 0.418 Views

- 19.0 Likes + 561 Dislikes + 33 Comments - 0.472 Aggregate Followers

Predictor Coef SE Coef T P

Constant 49.64 16.37 3.03 0.003

Ratings -11356 3537 -3.21 0.002

Budget -0.94185 0.01538 -61.24 0.000

Screens 13.08 15.63 0.84 0.404

Views -0.41780 0.03928 -10.64 0.000

Likes -19.02 14.21 -1.34 0.182

Dislikes 560.7 418.1 1.34 0.181

Comments 32.9 122.5 0.27 0.788

Aggregate Followers -0.47249 0.02077 -22.75 0.000

S = 0.000272439 R-Sq = 96.0% R-Sq(adj) = 95.8%

1. Regresi dengan data yang telah diatasi *Missing Value* dan *outlier*nya serta telah ditransformasikan

**Regression Analysis: Gross versus Ratings, Budget, ...**

The regression equation is

Gross = 43.8 - 9105 Ratings - 0.981 Budget + 12.8 Screens - 0.537 Views

+ 19.7 Likes - 39 Dislikes + 444 Comments - 0.489 Aggregate Followers

Predictor Coef SE Coef T P

Constant 43.79 18.43 2.38 0.018

Ratings -9105 3638 -2.50 0.013

Budget -0.98053 0.01659 -59.10 0.000

Screens 12.79 15.85 0.81 0.421

Views -0.53722 0.05337 -10.07 0.000

Likes 19.73 19.54 1.01 0.314

Dislikes -39.2 554.1 -0.07 0.944

Comments 444.0 185.8 2.39 0.018

Aggregate Followers -0.48931 0.02142 -22.85 0.000

S = 0.000302495 R-Sq = 96.1% R-Sq(adj) = 95.9%

Perbandingan kebaikan model regresi :

|  |  |  |
| --- | --- | --- |
|  |  | R-Sq |
| 1. | Missing Value | 61.9% |
| 2. | Missing Value dan Outlier | 65.3% |
| 3. | Missing Value dan Transformasi | 96.0% |
| 4. | Missing Value, Outlier dan Transformasi | 96.1% |

R squared merupakan angka yang berkisar antara 0 sampai 1 yang mengindikasikan besarnya kombinasi variabel independen secara bersama-sama mempengaruhi nilai variabel dependen. Semakin mendekati angka satu, model yang dikeluarkan oleh regresi tersebut akan semakin baik. Dari tabel perbandingan R-Sq diatas, data yang telah melalui pre-processing deteksi missing value memiliki nilai R-Sq yang lebih kecil dibandingkan dengan data yang telah melalui *pre-processing* deteksi *missing value* dan *outlier*. Begitu pula dengan data yang telah melalui *pre-processing* deteksi *missing value* dan *outlier* serta telah melewati proses transformasi akan memiliki nilai R-Sq yang lebih besar dibandingkan dengan data yang telah melalui *pre-processing* deteksi *missing value* dan telah melewati proses transformasi. Oleh karena itu, semakin baik (berkualitas) sebuah data maka akan menghasilkan model yang semakin baik juga.

1. Kesimpulan

Sebelum melakukan data mining perlu dilakukan *pre processing* data. Hal ini bertujuan untuk memastikan data yang akan diolah di data mining adalah data yang memiliki kualitas baik (tidak terdapat missing value, outlier, dll). Semakin baik (berkualitas) sebuah data maka akan menghasilkan model yang semakin baik juga.