4.26

1. Combine feature and salary to one file
2. Create a 10K data set
3. Use Weka, try linear regression
4. Try amazon machine learning

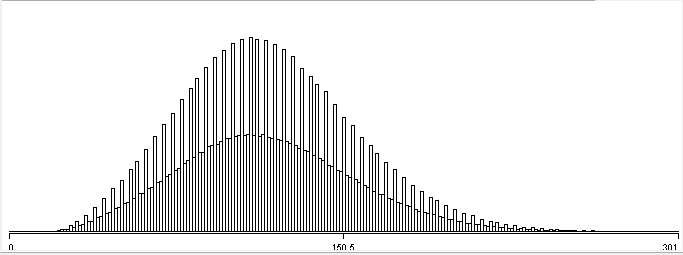


Figure salary distribution

* 1. use amazon machine learning

all data

**RMSE: 19.6411**

4.29

Use tableau analysis the data

Guess compId would effect salaries

Use panda select web ceo for compId=4

Create dummy features

4.30

Linear regression

Data: WEB CEO compId=4 df\_comp4\_dummies size : 322

scikit-learn. Cross-validation:

{'mean\_absolute\_error': -20.757251245699614,

'mean\_squared\_error': -627.41346121667323,

'r2': 0.40524298198516007}

---------------------------------------------------------------

Data: WEB CEO dummies size : 17907

scikit-learn. Cross-validation:

{'mean\_absolute\_error': -20.466599423737144,

'mean\_squared\_error': -603.04016314218734,

'r2': 0.46985733005148839}

{'mean\_absolute\_error': -20.466593915687053,

'mean\_squared\_error': -603.03940409704433,

'r2': 0.46985806655851547}

SVR(kernel=’rbf’)

{'mean absolute error': 22.147429765659943,

'mean squared error': 737.16208905894382,

'r2': 0.34567586840211684} one pass

{'mean\_absolute\_error': -21.939924703345323,

'mean\_squared\_error': -726.8560728746636,

'r2': 0.36122810231492841} cross

model = svm.SVR(kernel='linear')

{'mean absolute error': 20.338009979996492,

'mean squared error': 598.56007615776798,

'r2': 0.46469802835387364}

DecisionTreeRegressor

{'mean absolute error': 27.219709659408153,

'mean squared error': 1149.3148649626735,

'r2': -0.028167402641538608}

model = RandomForestRegressor(n\_estimators=10, max\_depth=None,min\_samples\_split=1, random\_state=0)

{'mean absolute error': 21.878114091648989,

'mean squared error': 719.37639282370071,

'r2': 0.36330824331465261}

model = AdaBoostRegressor(n\_estimators=100)

{'mean absolute error': 21.127839135497148,

'mean squared error': 643.24083868043476,

'r2': 0.43738942405836789}

model = GradientBoostingRegressor(n\_estimators=100, learning\_rate=0.1, max\_depth=1, random\_state=0, loss='ls')

{'mean absolute error': 20.902433490845084,

'mean squared error': 638.38377590889797,

'r2': 0.44265371820778621}

-----------------------------------------------

Weka Cross-validation

Correlation coefficient 0.6842

Mean absolute error 20.4981

Root mean squared error 24.6082

Relative absolute error 75.7041 %

Root relative squared error 72.925 %

Total Number of Instances 17907

All data

Linear

{'mean absolute error': 15.830457874095071,

'mean squared error': 383.89568530031977, RMSE = 19.59

'r2': 0.74410666813581905} one pass

{'mean\_absolute\_error': -15.844834015080949,

'mean\_squared\_error': -384.49101734938972,

'r2': 0.74350905079645147} cross validation

model = RandomForestRegressor(n\_estimators=200, max\_depth=None,min\_samples\_split=1, random\_state=0)

{'mean absolute error': 16.74735800590263,

'mean squared error': 442.93125389643939,

'r2': 0.70544859324101572}

Only web

{'mean\_absolute\_error': -16.312897650300908,

'mean\_squared\_error': -399.60268106019737,

'r2': 0.72564798037280176}