May 22nd, 2019: predicator variable and response variable

Get odd ratio, CV, predicated model, and data visualization.

Employee Attrition: get definition and investigate this article to write an introduction

We are examining the factors that to lead employee attrition.

<https://www.investopedia.com/terms/a/attrition.asp>

Summary of the Data:

1. we will examine the data set by looking at how many rows (subject) and how many columns (factors) we have.

* 1470 subjects and 35 variables

1. No missing data
2. **Data type**: 2 data type integer and factors
3. We are dealing with **imbalanced** **data**: 84% don’t leave their job (No Attrition) while 16% leave their job (Yes)

**Imbalanced data**: the number of observations in one class is significantly lower than the observation in the other class.

The predictive model developed using Conventional machine learning algorithms could be biased and inaccurate. Because the machine learning algorithm doesn’t take in consideration the distribution /the proportion or the balance of the classes.

<https://www.analyticsvidhya.com/blog/2017/03/imbalanced-classification-problem/>

* **how to handle imbalance datasets:**

1. **Data level approach**: resampling techniques: improving classification algorithms or balancing classes in the training data before providing data as input to the machine learning algorithm. The object is to increase the frequency of minority class or decrease the frequency of majority class.
2. **Algorithmic Ensemble Techniques**
3. Coding classification variables to numerical classification variables.

* **Attrition**

(Yes, No) → (1, 2)

* **Business Travel**

('Non-Travel', 'Travel\_Rarely', 'Travel\_Frequently') → (1, 2, 3)

* **Department**

('Human Resources', 'Research & Development', 'Sales') → (1, 2, 3)

* **Education Field**

('Human Resources', 'Life Sciences', 'Marketing', 'Medical', 'Technical Degree', 'Other') → (1, 2, 3, 4, 5)

* **Gender**

('Female', 'Male') → (1, 2)

* **Job Role**

('Healthcare Representative', 'Human Resources', 'Laboratory Technician', 'Manager', 'Manufacturing Director', 'Research Director', 'Research Scientist', 'Sales Executive', 'Sales Representative') → (1,2,3,4,5,6,7,8,9)

* **Marital Status**

('Single', 'Married', 'Divorced')

* **Over time**

('Yes','No') → (1,2)

1. Fitted logistic regression model

Look at this website for summary writing:

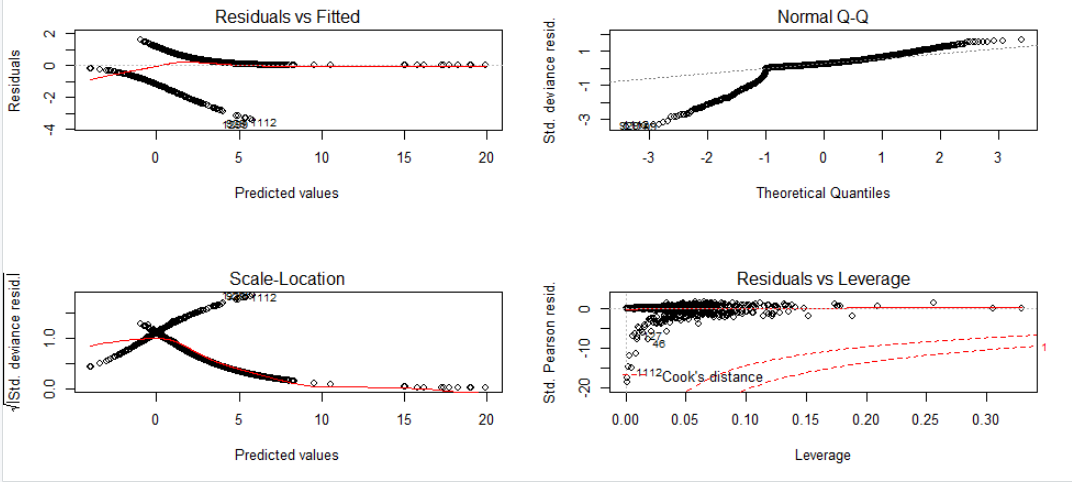
<https://stats.idre.ucla.edu/r/dae/logit-regression/>

* for the response Attrition without (Employee Count, Over18, and StandardHours) since these variables has the same answer for all subjects. And these variables effected the glm and mcDiagonase functions in R.
* Based on logistic model summary the significant variables are:

1. Age
2. Business Travel
3. Distance from home
4. Environment Satisfaction, Gender
5. Job Involvement
6. Job role 3: Laboratory Technician
7. Job Satisfaction
8. Marital Status 2 & 3: married and divorced
9. Number of companies Worked at
10. Overtime2
11. Relationship Satisfaction
12. Total Working Years
13. Training Time last year
14. Work Life Balance
15. Years at Company
16. Years in Current Role
17. Years Since Last Promotion
18. Years with Current Manager

* Diagnostic plots: look at this link for plot explanation.

<https://data.library.virginia.edu/diagnostic-plots/>



* VIF (Variance Inflation factor) from mc **mcDiagnose** Command:

We can conclude that there’s a multicollinearity situation for predicator variables that has a VIF value that exceeds 5 or 10:

Department 2 & 3: 44.05 / 43.95

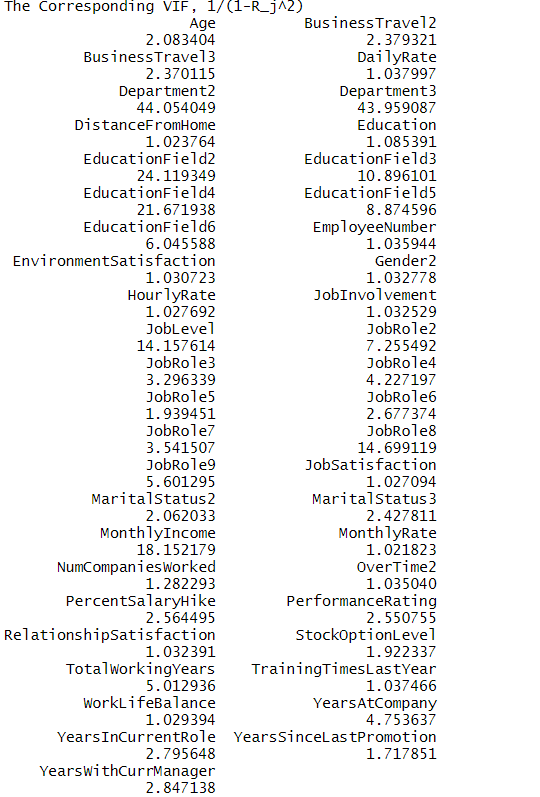
Education Field 2-6: 24.12 / 10.90 / 21.67 / 8.87 / 6.05

Job Level: 14.16

Job Role2 & 8: semi collinearity, the values are right on the edge collinearity. Values are (7.25/ 14.70)

Monthly income: 18.15

These variables were not significant in the generalized linear fitted model.



1. Model selection using

From the above 3-method for model selection, we should include:

Common Variables:

Intercept

Age

Business travel 2&3

Distance from home

Education Field>> best subset: 3 & 5

>> forward: 5

>> backward: 2, 4, &6

Environment Satisfaction

Gender2

Job Involvement

Job Role>> best subset: 2, 3, 8, &9

>> forward: 3, & 9

>>backward: 3, 8, &9

Marital Status: 2 &3

Number of Companies Worked

OverTime2

Relationship Satisfaction

Total Working Years

Training Times Last Year

Work Life Balance

Years at Company

Years in Current Role

Years Since Last Promotion

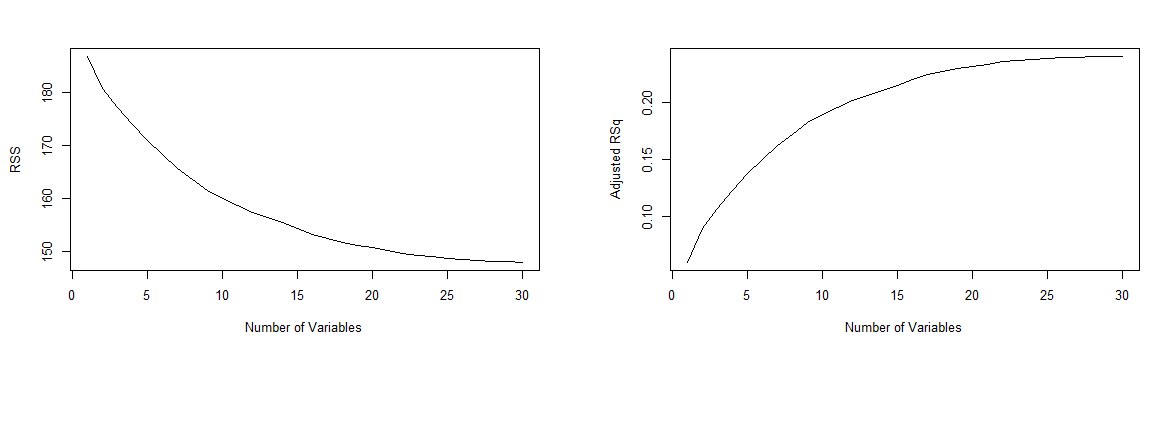
Years with Current Manager

1. **Best Subset Selection:** best described to have smallest RSS or equivalently to highest R2

Have to do model selection:

SML Ch6 slide 14 talks about AIC

* Plotting adjusted R2 and RSS
* R2 plot: we reach the highest R2 when the model has about 20 variables
* RSS plot: we also stay around 20 variables that we have the least number of variables with lower RSS values. Notice the big drop in RSS value when the model has 15 variables.



* Using which command in R:

which.max(bestsub.summary$adjr2)

[1] 30

> which.min(bestsub.summary$cp)

[1] 26

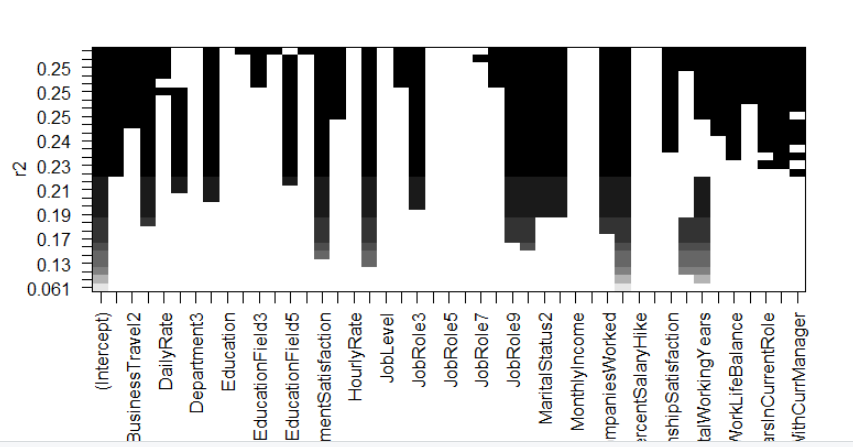
> which.min(bestsub.summary$bic)

[1] 17

* Now we’re going to look at what variables are included in each measure:

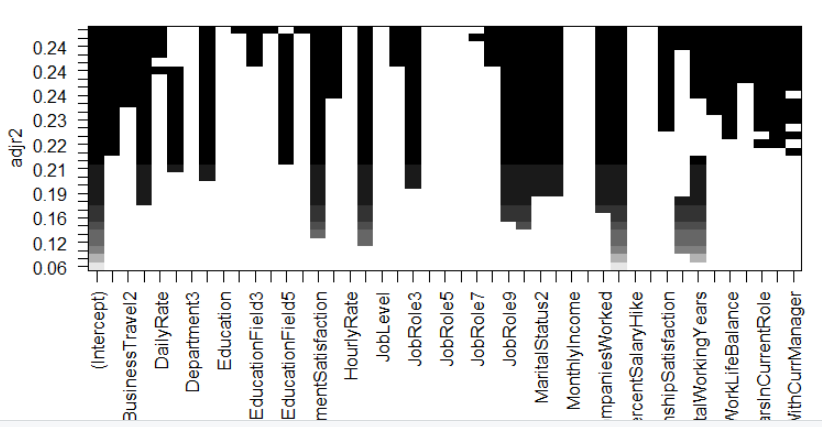
plot(bestsubset.fit, scale = "r2")

Black cubes mean specific variables are included and we’re looking for the highest R2



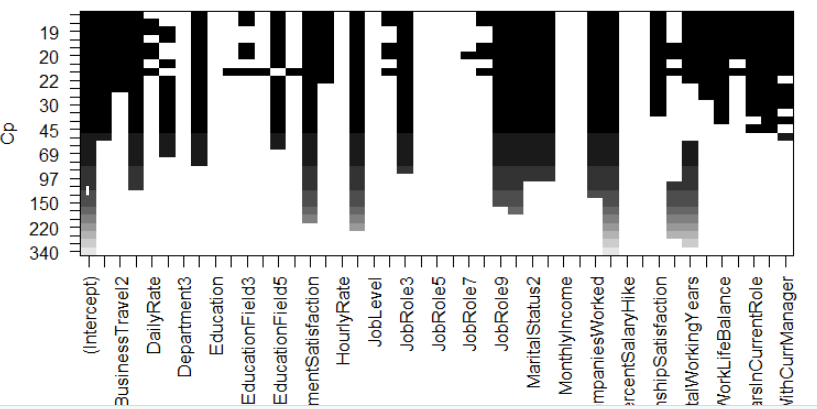
plot(bestsubset.fit, scale = "adjr2")

Black cubes mean specific variables are included and we’re looking for the highest adjusted-R2



plot(bestsubset.fit, scale = "Cp")

Black cubes mean specific variables are included and we’re looking for the lowest Cp



> coef(bestsubset.fit, which.min(bestsub.summary$cp))

(Intercept) Age BusinessTravel2

1.084512413 0.003490213 -0.065517959

BusinessTravel3 DistanceFromHome EducationField3

-0.154368308 -0.003457193 -0.051516555

EducationField5 EnvironmentSatisfaction Gender2

-0.101809713 0.040818917 -0.035497601

JobInvolvement JobRole2 JobRole3

0.058393564 -0.116951213 -0.111217450

JobRole8 JobRole9 JobSatisfaction

-0.051827960 -0.201866621 0.037824292

MaritalStatus2 MaritalStatus3 NumCompaniesWorked

0.115562063 0.137725678 -0.017001190

OverTime2 RelationshipSatisfaction TotalWorkingYears

0.211219745 0.022282302 0.004041715

TrainingTimesLastYear WorkLifeBalance YearsAtCompany

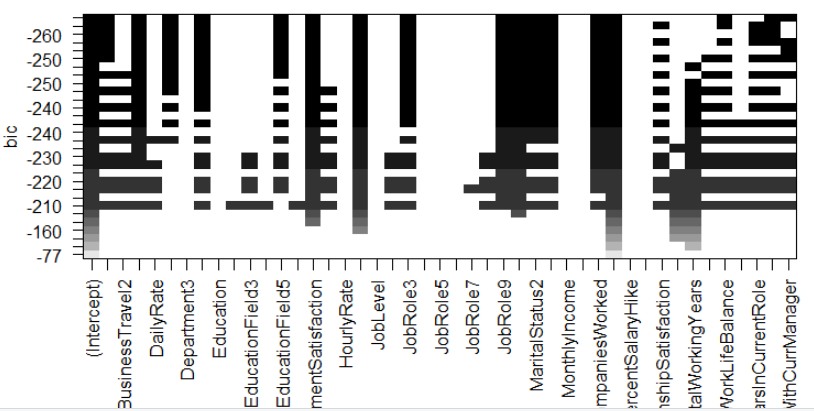
0.014206683 0.032829320 -0.005834251

YearsInCurrentRole YearsSinceLastPromotion YearsWithCurrManager

0.009555265 -0.011292028 0.009993043

plot(bestsubset.fit, scale = "bic")

Black cubes mean specific variables are included and we’re looking for the lowes BIC



We can call the coefficient of a set # of variables

coef(bestsubset.fit, 20)

1. **Forward Stepwise Selection:** best described to have smallest RSS or equivalently to highest R2

Using Cp as our statistical technique to measure the selection. We’re looking for the lowest Cp-value.

> coef(fwd.fit, which.min(summary(fwd.fit)$cp))

(Intercept) Age

9.905834e-01 3.391231e-03

BusinessTravel2 BusinessTravel3

-6.758817e-02 -1.569370e-01

DailyRate Department2

3.108053e-05 7.079285e-02

DistanceFromHome EducationField5

-3.541103e-03 -9.773182e-02

EnvironmentSatisfaction Gender2

4.055675e-02 -3.523420e-02

JobInvolvement JobRole3

5.861257e-02 -1.092935e-01

JobRole9 JobSatisfaction

-1.430211e-01 3.782097e-02

MaritalStatus2 MaritalStatus3

9.406633e-02 1.059331e-01

NumCompaniesWorked OverTime2

-1.736597e-02 2.110866e-01

RelationshipSatisfaction StockOptionLevel

2.211948e-02 1.793874e-02

TotalWorkingYears TrainingTimesLastYear

4.660636e-03 1.438780e-02

WorkLifeBalance YearsAtCompany

3.289145e-02 -5.168480e-03

YearsInCurrentRole YearsSinceLastPromotion

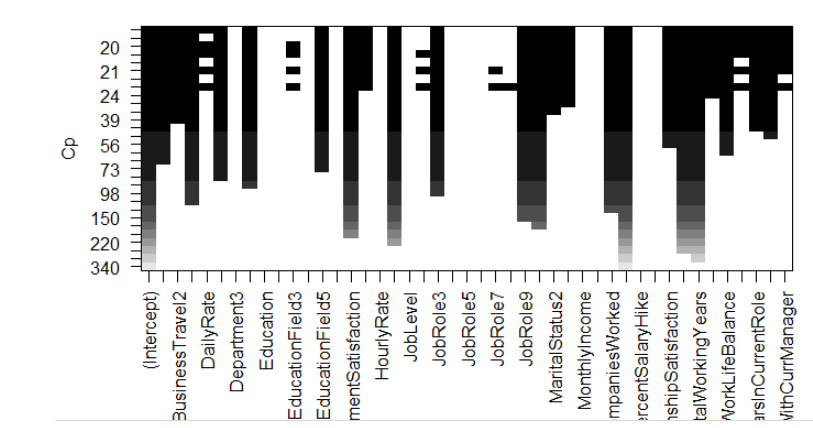
8.941654e-03 -1.112576e-02

YearsWithCurrManager

9.488499e-03

To better see the results of the forward selection using Cp statistics

plot(fwd.fit, scale = "Cp")



1. **Backward Elimination:** best described to have smallest RSS or equivalently to highest R2

Using Cp as our statistical technique to measure the selection. We’re looking for the lowest Cp-value.

> coef(bwd.fit, which.min(summary(bwd.fit)$cp))

(Intercept) Age

1.002441201 0.003537090

BusinessTravel2 BusinessTravel3

-0.065608705 -0.153825384

DistanceFromHome EducationField2

-0.003424638 0.074663173

EducationField4 EducationField6

0.088355961 0.093644973

EnvironmentSatisfaction Gender2

0.040684723 -0.035654912

JobInvolvement JobRole2

0.058348370 -0.084394947

JobRole3 JobRole8

-0.111060131 -0.039723324

JobRole9 JobSatisfaction

-0.193377748 0.037816881

MaritalStatus2 MaritalStatus3

0.117037014 0.138825816

NumCompaniesWorked OverTime2

-0.017102015 0.211135529

RelationshipSatisfaction TotalWorkingYears

0.022418202 0.004199780

TrainingTimesLastYear WorkLifeBalance

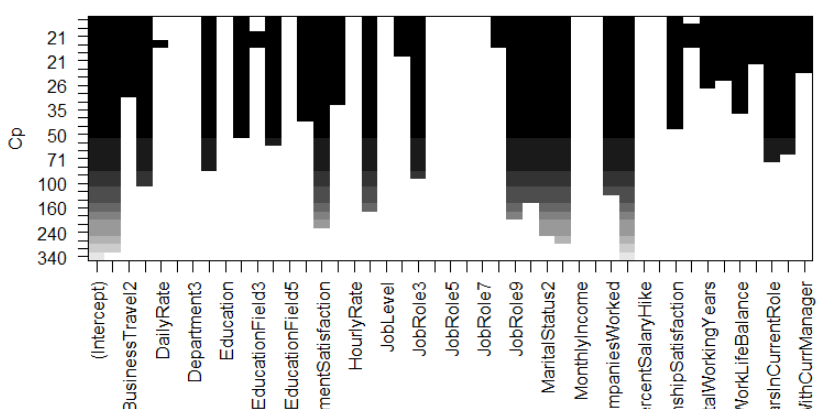
0.013485268 0.031875749

YearsAtCompany YearsInCurrentRole

-0.005662245 0.009426127

YearsSinceLastPromotion YearsWithCurrManager

-0.011513176 0.009846063



1. Cross-Validation: is a way to confirm our model selection.

http://www.science.smith.edu/~jcrouser/SDS293/labs/lab9-r.html

Regression using NN in OSCT folder Mahdi slide 15

Remember to do log ration for logistic model.

Plot qqnorm(residuals)

Plot diagnosis plots

Normality check for variables

The reputation of a company plays an important factor plays a part in:

1. Worker’s relationship with management
2. It effects employee’s confidence in the workplace and workload
3. Relations with customers. Training new employees to satisfy customer’s requires time. If employees are changed often customers take their business elsewhere
4. Hiring, training, and lost of production cost employer 150% departure employees

Source:

<https://work.chron.com/effects-employee-resignations-productivity-11269.html>