Progetto per l'esame di Statistica progredito

Federico Spatola

September 10, 2020

1 Introduzione

Il progetto è stato sviluppato utilizzando un Dataset disponibile sul sito internet www.kaggle.com. I dati sono stati raccolti per un esigenza reale. Si riferiscono a un' azienda che ha un numero di impiegati costante nel tempo, pari a quattro mila. Era noto che ogni anno circa il quindici per cento di loro lasciava la compagnia obbligandola a nuove assunzioni. Inoltre era evidente che questo meccanismo fosse poco virtuoso, di conseguenza è stata contattata un azienda che si occupa di analisi delle risorse umane [1] per comprendere il comportamento degli impiegati e quindi quali misure adottare affinchè negli anni futuri non avvengano così tanti licenziamenti.

Il dataset è costituito di cinque tabelle di dati. Le due tabelle *employee-survey-data.csv* e *manager-survey-data.csv* contengono dati relativi al livello di soddisfazione degli impiegati. La tabella *general-data.csv* contiene informazioni relative agli impiegati, ad esempio la loro età, la distanza tra la loro casa e il luogo di lavoro, l'impiego che hanno all'interno dell'azienda e una variabile, chiamata *Attrition*, che assume valore "No" se l'impiegato non ha lasciato l'azienda e "Sì" altrimenti. Le ultime due tabelle *in-time.csv* e *out-time.csv* danno informazioni riguardo ai tempi di login e logout nelle giornate di lavoro che vanno dal 01/01/2015 al 31/12/2015.

Nel seguito si è costruito un modello di regressione logistica in grado di stabilire per qualsiasi impiegato la probabilità che lui ha di lasciare l'azienda.

2 Preprocessing e trasformazione dei dati

Nello svolgere questo progetto, si è posto il problema di come utilizzare le due tabelle di dati relative ai tempi di login e logout. In particolare, c'era la necessità di trasformarle in variabili utilizzabili nel modello di regressione logistica. Dunque si è scelto di definire una nuova variabile *AvgWorkHrs* che, relativamente ad ogni impiegato, avesse come valore la media delle differenze, calcolate giorno per giorno, tra i tempi di logout e quelli di login. Ad esempio se un impiegato va a lavorare per tutto il periodo di osservazione alle nove di mattina e termina di lavorare alle diciassette allora a questo impiegato viene assegnato il numero otto come valore della nuova variabile.

Quindi si è effettuato il "merging", con cui si è ottenuta un unica tabella di dati. Sono state anche rimosse le righe che contenevano valori mancanti per qualche variabile. Piuttosto che eliminare tali righe, si è anche provato a sostituire i suddetti valori mancanti con dei valori medi ma questa scelta non ha dato risultati soddisfacenti. Si è cercato dei valori anomali come ad esempio anni di lavoro o età dei partecipanti non compatibili con la realtà ma non ne sono stati trovati.

Si è guardato poi alle correlazioni tra le variabili. Qui di seguito sono riportate le variabili che hanno un coefficiente di correlazione di Paerson maggiore di 0.3. In particolare la variabile "Attrition" non è fortemente correlata con nessuna delle altre.

	X1	X2	value
155	PercentSalaryHike	PerformanceRating	0.7739022
519	PerformanceRating	PercentSalaryHike	0.7739022
647	YearsWithCurrManager	YearsAtCompany	0.7686997
699	YearsAtCompany	YearsWithCurrManager	0.7686997
184	TotalWorkingYears	Age	0.6812131
574	Age	TotalWorkingYears	0.6812131
591	YearsAtCompany	TotalWorkingYears	0.6331943
643	TotalWorkingYears	YearsAtCompany	0.6331943
646	YearsSinceLastPromotion	YearsAtCompany	0.6193416
672	YearsAtCompany	YearsSinceLastPromotion	0.6193416
674	YearsWithCurrManager	YearsSinceLastPromotion	0.5100808
700	YearsSinceLastPromotion	YearsWithCurrManager	0.5100808
593	YearsWithCurrManager	TotalWorkingYears	0.4633693
697	TotalWorkingYears	YearsWithCurrManager	0.4633693
592	YearsSinceLastPromotion	TotalWorkingYears	0.4070773
670	TotalWorkingYears	YearsSinceLastPromotion	0.4070773
186	YearsAtCompany	Age	0.3143842
628	Age	YearsAtCompany	0.3143842
181	NumCompaniesWorked	Age	0.3037089
493	Age	NumCompaniesWorked	0.3037089

3 Visualizzazione dei dati

Prima di costruire il modello è ragionevole ottenere una migliore comprensione dei dati a disposizione. L'utilizzo di strumenti come gli istogrammi, i bar-plots e i box-plots aiuta a mettere in evidenza tendenze e valori anomali. Si è scelto di riportare le visualizzazioni di quelle variabili che si ritengono di maggiore interesse. La Figura 1 mostra che i Single hanno una maggiore tendenza a licenziarsi, rispetto alle persone che risultano sposate o divorziate. Si può dire anche che gli impiegati di alcuni Dipartimenti hanno una maggiore tendenza a lasciare il posto di lavoro rispetto a lavoratori che sono impiegati in altri settori. Infatti nel Dipartimenti di Ricerca e sviluppo e in quello delle Vendite il tasso degli impiegati che cambiano lavoro è inferiore al venti per cento, mentre circa il trenta per cento degli impiegati nelle risorse umane sceglie di licenziarsi. Analogamente tra quelli che non sono soddisfatti dell'ambiente di lavoro c'è una maggiore tendenza al licenziamento. Invece il tasso di lavoratori che lascia il posto di lavoro è pressochè lo stesso per tutti i livelli della variabile JobLevel.

Figure 1: Barplots

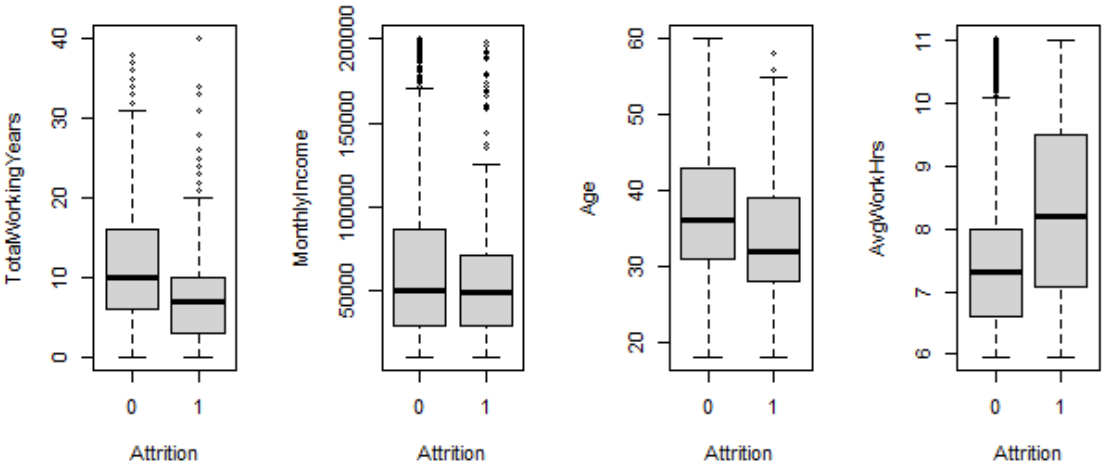

Figure 2: Boxplots

Nella figura 2 si vede che i lavoratori che lasciano il posto di lavoro hanno tendenzialmente meno anni di lavoro. Ad esempio si può notare che il settantacinque per cento dei lavoratori che smettono di lavorare ha meno di dieci anni di lavoro mentre il cinquanta per cento dei lavoratori che continuano a lavorare nell'azienda lavora già da più di dieci anni. La mediana degli stipendi relativi ai lavoratori che scelgono di licenziarsi è la stessa di quella dei lavoratori che continuano a lavorare per l'azienda. Dunque sembrerebbe che lo stipendio non è tra i fattori che influenzano maggiormente il lavoratore ad andare via. Si può dire anche che la mediana dell'età è più bassa per i lavoratori che scelgono di licenziarsi. Inoltre i lavoratori che si licenziano lavorano più ore. I boxplots ci informano anche della presenza di outliers. Ad esempio vengono segnalati come outliers le età dei lavoratori che si licenziano e che hanno più di cinquantacinque anni. Si tratta in ogni caso di valori plausabili quindi si è scelto di non eliminarli dall'insieme di dati.

4 Modello di regressione logistica

La regressione logistica in R è implementata con la funzione *glm()*, utilizzando l'opzione family=binomial. Il modello di regressione logistica è un modello binomiale che utilizza la funzione di ripartizione della variabile aleatoria logistica come legame, che è il legame canonico se l'ipotesi distributiva è appunto binomiale. In R, se non è specificata la funzione legame, l'implementazione del modello lineare generalizzato utilizza la funzione legame canonica relativa all' ipotesi distributiva, dunque nella funzione *glm()* non è stato necessario specificare l'opzione *link=logit*.

```
glm(formula = Attrition ~ ., family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8160  -0.5713  -0.3450  -0.1618   3.6396

Coefficients:
                                 Estimate Std. Error z value Pr(>|z|)
(Intercept)                     7.090e-02  7.757e-01   0.091  0.92718
EmployeeID                     -5.361e-06  3.703e-05  -0.145  0.88488
EnvironmentSatisfaction        -3.823e-01  4.344e-02  -8.802  < 2e-16 ***
JobSatisfaction                -3.569e-01  4.345e-02  -8.213  < 2e-16 ***
WorkLifeBalance                -3.493e-01  6.561e-02  -5.324 1.02e-07 ***
JobInvolvement                 -9.194e-02  6.597e-02  -1.394  0.16339
PerformanceRating              -5.363e-03  2.050e-01  -0.026  0.97913
Age                            -3.324e-02  7.546e-03  -4.405 1.06e-05 ***
BusinessTravelTravel_Frequently 1.414e+00  2.104e-01   6.718 1.84e-11 ***
BusinessTravelTravel_Rarely     6.243e-01  1.962e-01   3.183  0.00146 **
DepartmentResearch & Development -5.883e-01  2.849e-01  -2.065  0.03894 *
DepartmentSales                -5.864e-01  2.982e-01  -1.966  0.04926 *
DistanceFromHome               -3.143e-03  6.019e-03  -0.522  0.60154
Education                      -6.006e-02  4.725e-02  -1.271  0.20365
EducationFieldLife Sciences    -5.942e-01  3.903e-01  -1.522  0.12790
EducationFieldMarketing        -8.486e-01  4.266e-01  -1.989  0.04668 *
EducationFieldMedical          -6.663e-01  3.900e-01  -1.709  0.08751 .
EducationFieldOther            -9.409e-01  4.367e-01  -2.154  0.03120 *
EducationFieldTechnical Degree -9.316e-01  4.195e-01  -2.221  0.02637 *
GenderMale                      6.644e-02  9.840e-02   0.675  0.49955
JobLevel                       -6.360e-02  4.386e-02  -1.450  0.14704
JobRoleHuman Resources         -1.335e-01  3.121e-01  -0.428  0.66885
JobRoleLaboratory Technician    1.219e-01  1.999e-01   0.610  0.54202
JobRoleManager                 -3.348e-01  2.561e-01  -1.307  0.19108
JobRoleManufacturing Director  -5.403e-01  2.374e-01  -2.276  0.02283 *
JobRoleResearch Director        5.990e-01  2.450e-01   2.446  0.01446 *
JobRoleResearch Scientist       1.966e-01  1.938e-01   1.015  0.31031
JobRoleSales Executive          3.130e-01  1.920e-01   1.630  0.10309
JobRoleSales Representative    -1.305e-01  2.580e-01  -0.506  0.61302
MaritalStatusMarried            2.914e-01  1.431e-01   2.037  0.04165 *
MaritalStatusSingle             1.177e+00  1.430e-01   8.229  < 2e-16 ***
MonthlyIncome                  -1.104e-06  1.040e-06  -1.061  0.28854
NumCompaniesWorked              1.481e-01  2.049e-02   7.230 4.85e-13 ***
PercentSalaryHike               1.356e-02  2.061e-02   0.658  0.51054
StockOptionLevel               -7.806e-02  5.686e-02  -1.373  0.16979
TotalWorkingYears              -8.161e-02  1.365e-02  -5.980 2.23e-09 ***
TrainingTimesLastYear          -1.556e-01  3.818e-02  -4.075 4.60e-05 ***
YearsAtCompany                  3.800e-02  1.966e-02   1.933  0.05327 .
YearsSinceLastPromotion         1.643e-01  2.212e-02   7.429 1.09e-13 ***
YearsWithCurrManager           -1.772e-01  2.478e-02  -7.152 8.57e-13 ***
AvgWorkhrs                      4.414e-01  3.417e-02  12.918  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3804.3  on 4299  degrees of freedom
Residual deviance: 2946.7  on 4259  degrees of freedom
AIC: 3028.7

Number of Fisher Scoring iterations: 6
```

I segni dei coefficienti stimati sono quelli che ci si attendeva. In particolare, si aspettavano segni negativi per i coefficienti delle variabili "TotalWorkingYears","Age" e segno positivo per "AvgWorkHrs". I p-value rivelano che per alcune variabili non si rigetta l'ipotesi nulla dei loro coefficienti. Dunque il campione non ha provveduto sufficiente evidenza per concludere che esiste un effetto di queste variabili sulla media della variabile dipendente.
Si è deciso di utilizzare il criterio d'informazione di Akaike per migliorare il "fitting" dell modello. Per fare ciò in R si è utilizzata la funzione *stepAIC()*.

```
glm(formula = Attrition ~ EnvironmentSatisfaction + JobSatisfaction +
    WorkLifeBalance + JobInvolvement + Age + BusinessTravel +
    Department + JobLevel + JobRole + MaritalStatus + NumCompaniesWorked +
    TotalWorkingYears + TrainingTimesLastYear + YearsAtCompany +
    YearsSinceLastPromotion + YearsWithCurrManager + AvgWorkhrs,
    family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8091  -0.5707  -0.3496  -0.1621   3.7315

Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                    -0.475778   0.568546  -0.837  0.40269
EnvironmentSatisfaction        -0.376551   0.043203  -8.716  < 2e-16 ***
JobSatisfaction                -0.357507   0.043094  -8.296  < 2e-16 ***
WorkLifeBalance                -0.334991   0.065057  -5.149 2.62e-07 ***
JobInvolvement                 -0.097694   0.065480  -1.492  0.13571
Age                            -0.033209   0.007387  -4.496 6.94e-06 ***
BusinessTravelTravel_Frequently  1.414398   0.209251   6.759 1.39e-11 ***
BusinessTravelTravel_Rarely      0.629559   0.194786   3.232  0.00123 **
DepartmentResearch & Development -0.906611   0.196431  -4.615 3.92e-06 ***
DepartmentSales                -0.973202   0.206348  -4.716 2.40e-06 ***
JobLevel                       -0.071798   0.043347  -1.656  0.09765 .
JobRoleHuman Resources         -0.175704   0.306890  -0.573  0.56696
JobRoleLaboratory Technician    0.117955   0.197308   0.598  0.54996
JobRoleManager                 -0.316444   0.252936  -1.251  0.21090
JobRoleManufacturing Director  -0.554546   0.234244  -2.367  0.01791 *
JobRoleResearch Director        0.563134   0.242455   2.323  0.02020 *
JobRoleResearch Scientist       0.184530   0.191199   0.965  0.33448
JobRoleSales Executive          0.270664   0.189379   1.429  0.15294
JobRoleSales Representative    -0.128102   0.255468  -0.501  0.61606
MaritalStatusMarried            0.273002   0.141048   1.936  0.05293 .
MaritalStatusSingle             1.182403   0.140805   8.397  < 2e-16 ***
NumCompaniesWorked              0.147256   0.020225   7.281 3.32e-13 ***
TotalWorkingYears              -0.082831   0.013596  -6.092 1.11e-09 ***
TrainingTimesLastYear          -0.149981   0.037766  -3.971 7.15e-05 ***
YearsAtCompany                  0.036396   0.019549   1.862  0.06263 .
YearsSinceLastPromotion         0.162297   0.021994   7.379 1.60e-13 ***
YearsWithCurrManager           -0.175342   0.024662  -7.110 1.16e-12 ***
AvgWorkhrs                      0.451457   0.033922  13.309  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3804.3  on 4299  degrees of freedom
Residual deviance: 2962.7  on 4272  degrees of freedom
AIC: 3018.7

Number of Fisher Scoring iterations: 6
```

Il nuovo modello ha ovviamente un AIC più piccolo di quello precedente e quindi risulta da questo punto di vista migliore. Inoltre è diminuito il numero delle variabili per cui non si rifiuta l'ipotesi di nullità del coefficiente di regressione. Le variabili per cui non si rifiuta tale ipotesi nulla sono alcune delle variabili "Dummies" che sono state ricavate da variabili categoriali, ad esempio *JobRoleSales Representative* e *MaritalStatusMarried*. Tuttavia il nuovo output della funzione *glm()* non mostra una più bassa devianza dei residui. Invece il numero di iterazioni dell'algoritmo di Fisher Scoring impiegato nella stima dei coefficienti, che era uguale a sei, è rimasto inalterato.
Questo modello potrebbe risultare molto utile all'azienda che ha messo a disposizioni questi dati. Si puo' avere un indicazione di quali impiegati potrebbero lascire l'azienda nei prossimi anni e in taluni casi provare a trattenerli aumentandogli gli stipendi o attuando altri tipi di misure. Di seguito viene mostrato l'esempio di un impiegato per cui il modello calcola che c'è un rischio molto alto di licenziamento.

```
impiegato=data.frame(EnvironmentSatisfaction = 1,
                     Department="Sales",
                     Gender="Male",
                     JobSatisfaction  = 1,
                     WorkLifeBalance   =  1,
                     PerformanceRating  =   1,
                     PercentSalaryHike =11,
                     Age  =  20,
                     BusinessTravel="Non-Travel",
                     JobLevel=1,
                     DistanceFromHome =   22,
                     EducationField ="Life Sciences",
                     JobRole="Sales Executive",
                     MaritalStatus="Single"  ,
                     NumCompaniesWorked =0  ,
                     TotalWorkingYears  =0  ,
                     YearsWithCurrManager  =0,
                     JobInvolvement=0,
                     YearsAtCompany=0,
                     YearsSinceLastPromotion=0,
                     TrainingTimesLastYear=0,
                     AvgWorkhrs  =12 )
predict(model2,newdata=impiegato, type="response")
       1
0.9738278
```

5 Conclusioni

Il modello sembra lavorare in modo soddisfacente, tuttavia si potrebbe agire in diversi modi per migliorarlo. Dalle due tabelle *in-time.csv* e *out-time.csv* si potrebbero creare altre variabili, ad esempio si può definire una variabile orario medio di uscita dal lavoro che potrebbe avere un effetto sulla variabile dipendente ed essere comunque non troppo correlata con $AvgWorkHrs$. Si può provare a vedere se ci sono effetti non lineari o usare altre funzioni legame come ad esempio il legame probit o quello log-log complementare. Dopo il "merging", si potrebbe anche dividere le osservazioni in training dataset e testing dataset e creare un modello di classificazione. Quindi si costruirebbe un modello di regressione logistica utilizzando il primo insieme di dati e poi se ne verificherebbe la validità impostando un valore cut-off e guardando all'errore di classificazione sul secondo insieme di dati.

References

[1] Human Resource Analytics (2020):
https://www.valamis.com/hub/hr-analytics#:~:text=HR%20analytics%20is%20the%20process,improve%20an%20organization's%20workforce%20performance.&text=HR%20analytics%20provides%20data%2Dbacked,more%20effectively%20for%20the%20future.