Machine Learning Analysis

By:
Anthony Stachowitz

1. Which machine learning technique will you use?

Logistical regression is a way to fit a regression curve where y = f(x). This model is used to predict y given a set of predictors x. The predictors in this model can be continuous, categorical or a mix of both. I am using categorical predictors for my model.

I am using a binominal logistic regression because y is binary, it can only be 1 or 0. Given a set of attributes for each person in the census such as education level, occupation, marital status and class of work, the algorithm should decide whether the person makes less than $50,000 (1) or more than $50,000 (0).

1. How do you frame your main question as a machine learning problem? Is it a supervised or unsupervised problem? If it is supervised, is it a regression or a classification?

Binominal logistical regression is a supervised machine learning problem. We have both, the independent variables, and the dependent variables. The independent variables in my model are class of work, education level, occupation, and marital status. The dependent variable is whether the income is over $50,000 or under $50,000.

The model is considered a classification predictive model. There are two classes that are being predicted, over $50,000 and under $50,000.

1. What are the main features (also called independent variables or predictors) that you'll use?

Each independent variable has a different number of features.

Marital status includes 7 classifications. They are Married to a civilian, Married to an armed forces spouse, Married but not together, Divorced, Separated, Widowed, or Never married.

Education includes 16 classifications. They are Preschool, 1st to 4th grade, 5th to 6th grade, 7th to 8th grade, 9th grade, 10th grade, 11th grade, 12th grade, High school graduate, Some college, Vocational associates degree, Academic associates degree, Professional school, Bachelor's degree, Master's degree, and Doctorate degree.

Occupation includes 14 classifications. They are Administrational-clerical, Armed Forces, Craft repair, Executive-managerial, Farming-fishing, Handlers-cleaners, Machine operator-inspector, Other service, Private house service, Professional-specialty, Protective service, Sales, Tech-support, Transportation-moving.

Working class includes 8 classifications. They are Federal government, Local government, Never worked, Private, Self-employed incorporation, Self-employed not incorporated, State government, Without pay.

```
> model <- glm(LessThen_50 ~ WorkClass + Education + MaritalStatus + Occupati
on,family=binomial(link='logit'),data=income_data_f)
> summary(model)

Call:
glm(formula = LessThen_50 ~ WorkClass + Education + MaritalStatus + 
    Occupation, family = binomial(link = "logit"), data = income_data_f)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-3.4495   0.0793   0.2575   0.5845   2.4744  

Coefficients: (1 not defined because of singularities)
                                    Estimate Std. Error z value Pr(>|z|)    
(Intercept)                          4.27902    0.17282  24.760  < 2e-16 ***
WorkClassFederal-gov                -1.18550    0.13953  -8.496  < 2e-16 ***
WorkClassLocal-gov                  -0.46962    0.12701  -3.697 0.000218 ***
WorkClassNever-worked               11.11977  291.15371   0.038 0.969535    
WorkClassPrivate                    -0.63769    0.11185  -5.701 1.19e-08 ***
WorkClassSelf-emp-inc               -1.17868    0.13480  -8.744  < 2e-16 ***
WorkClassSelf-emp-not-inc           -0.37428    0.12368  -3.026 0.002477 ** 
WorkClassState-gov                  -0.20785    0.13854  -1.500 0.133535    
WorkClassWithout-pay                12.10596  204.23048   0.059 0.952732    
Education11th                        0.06519    0.19738   0.330 0.741206    
Education12th                       -0.37086    0.24104  -1.539 0.123912    
Education1st-4th                     0.77263    0.44883   1.721 0.085175 .  
Education5th-6th                     0.41805    0.29798   1.403 0.160637    
Education7th-8th                     0.42443    0.21885   1.939 0.052455 .  
Education9th                         0.36928    0.24855   1.486 0.137350    
EducationAssoc-acdm                 -1.24185    0.16338  -7.601 2.94e-14 ***
EducationAssoc-voc                  -1.24731    0.15676  -7.957 1.76e-15 ***
EducationBachelors                  -1.91459    0.14578 -13.134  < 2e-16 ***
EducationDoctorate                  -3.30344    0.19957 -16.553  < 2e-16 ***
EducationHS-grad                    -0.73860    0.14247  -5.184 2.17e-07 ***
EducationMasters                    -2.42527    0.15565 -15.582  < 2e-16 ***
EducationPreschool                  11.43366  106.24244   0.108 0.914298    
EducationProf-school                -3.13069    0.18512 -16.912  < 2e-16 ***
```

```
EducationSome-college               -1.04403    0.14429  -7.236 4.63e-13 ***
MaritalStatusMarried-AF-spouse      -2.39984    0.46428  -5.169 2.35e-07 ***
MaritalStatusMarried-civ-spouse     -2.08355    0.05675 -36.717  < 2e-16 ***
MaritalStatusMarried-spouse-absent   0.11999    0.19953   0.601 0.547608    
MaritalStatusNever-married           0.85325    0.07153  11.928  < 2e-16 ***
MaritalStatusSeparated               0.27591    0.14502   1.903 0.057100 .  
MaritalStatusWidowed                -0.12532    0.13260  -0.945 0.344593    
OccupationAdm-clerical               0.09969    0.08932   1.116 0.264401    
OccupationArmed-Forces               1.17453    1.26734   0.927 0.354048    
OccupationCraft-repair              -0.06954    0.07951  -0.875 0.381822    
OccupationExec-managerial           -0.89943    0.08113 -11.086  < 2e-16 ***
OccupationFarming-fishing            0.67258    0.12987   5.179 2.23e-07 ***
OccupationHandlers-cleaners          0.84740    0.13713   6.180 6.42e-10 ***
OccupationMachine-op-inspct          0.37896    0.09988   3.794 0.000148 ***
OccupationOther-service              1.02791    0.11530   8.915  < 2e-16 ***
OccupationPriv-house-serv            2.45531    1.01952   2.408 0.016027 *  
OccupationProf-specialty            -0.51055    0.08648  -5.904 3.55e-09 ***
OccupationProtective-serv           -0.59115    0.12289  -4.811 1.51e-06 ***
OccupationSales                     -0.33671    0.08352  -4.031 5.54e-05 ***
OccupationTech-support              -0.54547    0.11129  -4.901 9.51e-07 ***
OccupationTransport-moving                NA         NA      NA       NA    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 35948  on 32560  degrees of freedom
Residual deviance: 23861  on 32518  degrees of freedom
AIC: 23947

Number of Fisher Scoring iterations: 13
```