PREDICTING STUDENT DAYLY ALCOHOL CONSUMPTION

By Pavel Dudin, Charlotte 2016

Introduction

Why underage drinking is a risky business

- Accidents and Injuries
- Alcohol Poisoning
- Brain Development
- Mental Health
- Aggression and Violence
- Liver Damage

Goal

The final goal is to find the best predicting model and correlation between alcohol consumption over the week.

Weekly consumption more significant weekend.

Data set

- Date (2005-2006)
- 1044 students Two school in Portugal
- Students Alcohol Consumption
- https://archive.ics.uci.edu/ml/datasets/STUDENT+ALCOHOL+CONSUMPTION
- 33 variables

Data processing

Models used:

1. Linear model
2. Logistic prediction model
3. Multinomial model
4. CVM model

Linear model

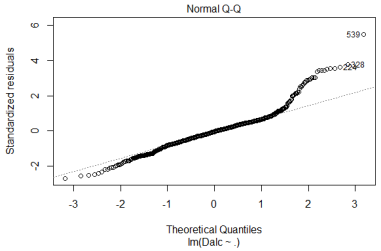
- Main factors affected Daily Alcohol Consumption

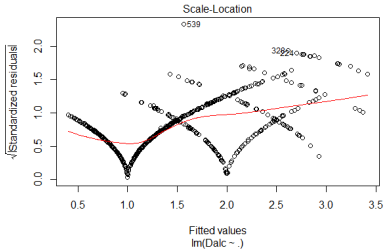
sexM	0.183566	0.059005	3.111	0.001950 **
Medu	0.127691	0.039199	3.258	0.001185 **
Fedu	-0.065954	0.033547	-1.966	0.049737 *
Mjobhealth	-0.409413	0.136727	-2.994	0.002859 **
Fjobother	-0.327085	0.130334	-2.510	0.012339 *
reasonother	0.345565	0.102068	3.386	0.000755 ***
schoolsupyes	0.144666	0.079771	1.814	0.070232 .
nurseryyes	-0.116294	0.067679	-1.718	0.086235 .
freetime	0.088547	0.028698	3.085	0.002122 **
Walc	0.410729	0.024550	16.730	< 2e-16 ***

Logistic prediction model

	FALSE	TRUE
1	467	5
2	111	11
3	22	18
4	6	10
5	1	15

with threshold 2.5



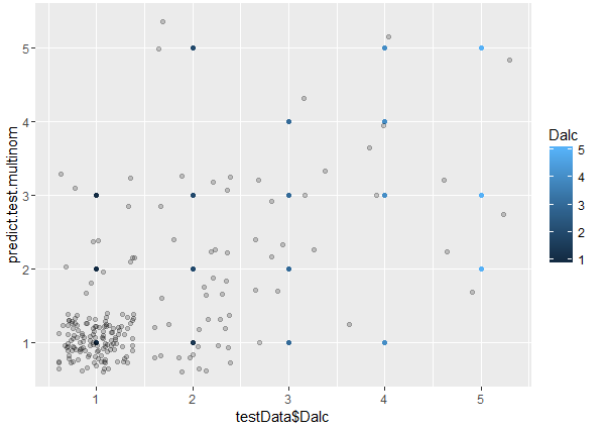


Multinomial Regression Model

The confusion matrix looks like this table predict.test.multinom

	1	2	3	4	5
1	129	9	4	0	0
2	20	10	5	0	2
3	1	6	4	1	0
4	1	0	1	2	1
5	0	2	2	0	1

misclassification error 27.3% ,low



SVM

```
Confusion Matrix and Statistics

          Reference
Prediction   1   2   3   4   5
         1 142   0   0   0   0
         2  22  15   0   0   0
         3   5   0   7   0   0
         4   1   0   0   4   0
         5   4   0   0   0   1

Overall Statistics

               Accuracy : 0.8408
                 95% CI : (0.7827, 0.8885)
    No Information Rate : 0.8657
    P-Value [Acc > NIR] : 0.8712

                  Kappa : 0.572
 Mcnemar's Test P-Value : NA
```

SVM with cross validation in R using caret

```
ctrl <- trainControl(method = "repeatedcv", repeats = 10)#setting up control

set.seed(1500)

mod <- train(Dalc ~ sex+ age+famsize+Pstatus+ Medu+Fedu +
             +                   studytime +failures+ schoolsup+ activities+
higher +romantic
              +                  famrel+freetime+goout, data=trainData, method
= "svmLinear", trControl = ctrl)
```

```
  RMSE         Rsquared
  0.9186448    0.09766363
```

Conclusion

- As a result, we determined that the best model to predict daily alcohol consumption for student population is SVM prediction model with accuracy of 0.84 in comparison with multinomial prediction model of 0.725. After cross validation using caret package RMSE equals 0.91 .
- Simple logistic regression gave us the most influential factors affected daily alcohol consumption. The key factors might be changed to decrease drinking are free time (positive correlation) and school support (positive correlation). Parents jobs and their educational level have a high impact although they are pretty stable and cannot be changed.