HW 7

Eli Sakov

Data Analytics

5.1.2015

**Bank Dataset**

Summary Statistics

age job marital education default

Min. :18.00 blue-collar:9732 divorced: 5207 primary : 6851 no :44396

1st Qu.:33.00 management :9458 married :27214 secondary:23202 yes: 815

Median :39.00 technician :7597 single :12790 tertiary :13301

Mean :40.94 admin. :5171 unknown : 1857

3rd Qu.:48.00 services :4154

Max. :95.00 retired :2264

(Other) :6835

balance housing loan contact day

Min. : -8019 no :20081 no :37967 cellular :29285 Min. : 1.00

1st Qu.: 72 yes:25130 yes: 7244 telephone: 2906 1st Qu.: 8.00

Median : 448 unknown :13020 Median :16.00

Mean : 1362 Mean :15.81

3rd Qu.: 1428 3rd Qu.:21.00

Max. :102127 Max. :31.00

month duration campaign pdays previous

may :13766 Min. : 0.0 Min. : 1.000 Min. : -1.0 Min. : 0.0000

jul : 6895 1st Qu.: 103.0 1st Qu.: 1.000 1st Qu.: -1.0 1st Qu.: 0.0000

aug : 6247 Median : 180.0 Median : 2.000 Median : -1.0 Median : 0.0000

jun : 5341 Mean : 258.2 Mean : 2.764 Mean : 40.2 Mean : 0.5803

nov : 3970 3rd Qu.: 319.0 3rd Qu.: 3.000 3rd Qu.: -1.0 3rd Qu.: 0.0000

apr : 2932 Max. :4918.0 Max. :63.000 Max. :871.0 Max. :275.0000

(Other): 6060

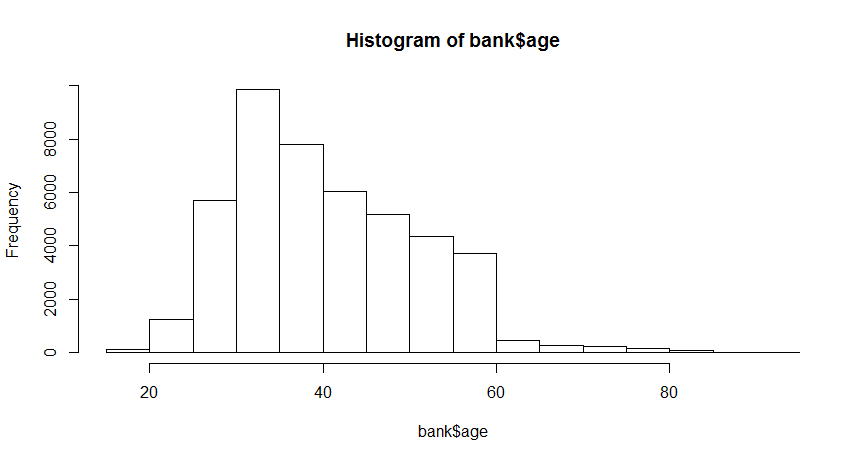
poutcome y

failure: 4901 no :39922

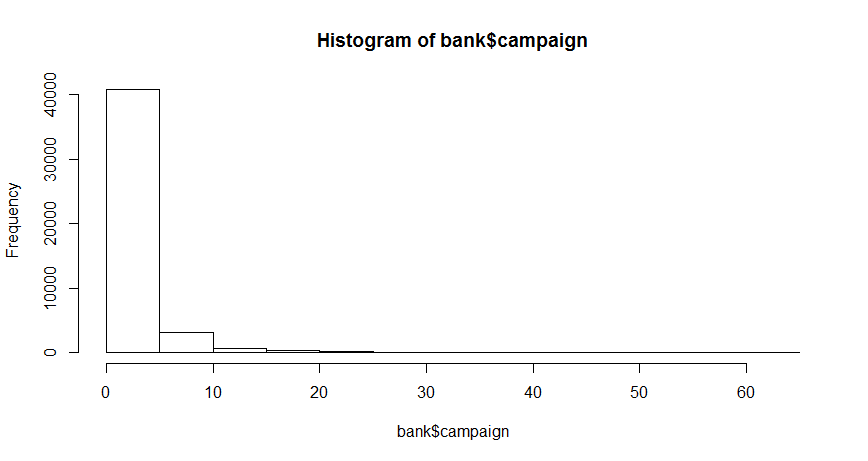
other : 1840 yes: 5289

success: 1511

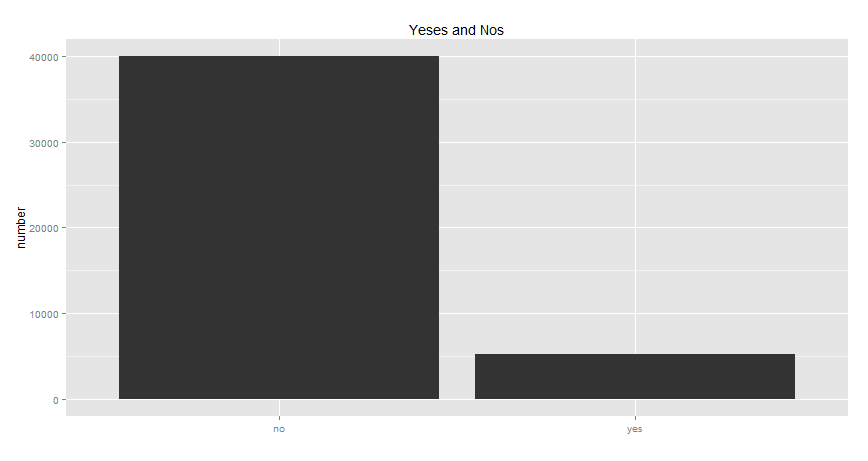
unknown:36959



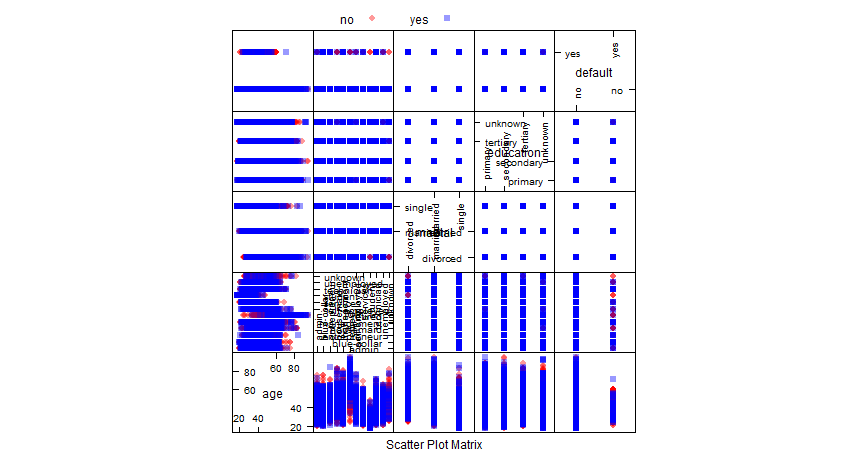
This distribution is skewed.



This is not normally distributed.



It appears that most responses are nos.



This is a feature plot matrix for 5 arbitrary features, since doing these for every variable would take a long time. It suggests weak features in the dataset, guiding me towards boosting. No transformations or cleaning on the dataset was performed since the library `caret` takes care of that. `caret` also creates the necessary factor variables. I expect to need a strong model to get good results from such weak features.

First I tried a boosted logistic regression on all the features, because logistic regressions are simple, good places to start. The library used handles validation, optimization, and tuning. I use 5-times by 10-fold cross validation.

Boosted Logistic Regression

45211 samples

16 predictor

2 classes: 'no', 'yes'

No pre-processing

Resampling: Cross-Validated (10 fold)

Summary of sample sizes: 40691, 40690, 40690, 40690, 40690, 40690, ...

Resampling results across tuning parameters:

nIter Accuracy Kappa Accuracy SD Kappa SD

11 0.8910221 0.3495045 0.002127892 0.06349747

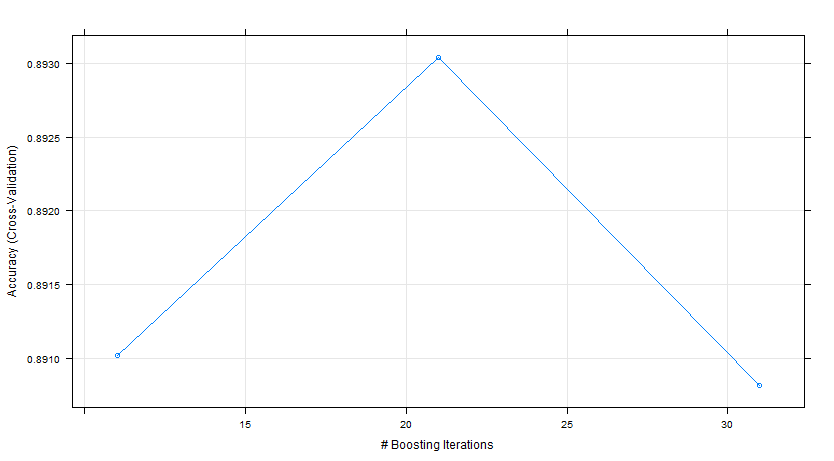
21 0.8930349 0.3064934 0.003926288 0.03694535

31 0.8908231 0.3697104 0.002266508 0.01102856

Accuracy was used to select the optimal model using the largest value.

The final value used for the model was nIter = 21.

Boosting Results



The model selected has an accuracy of 89%, which is approximately the percent of nos. The kappa statistic indicates fair performance. The tuning picked 21 iterations using 10-fold cross-validation. This is not a bad model

I then try a random forest because random forests are pretty good, in general.

Random Forest

45211 samples

16 predictor

2 classes: 'no', 'yes'

No pre-processing

Resampling: Cross-Validated (5 fold)

Summary of sample sizes: 36169, 36168, 36169, 36168, 36170

Resampling results across tuning parameters:

mtry Accuracy Kappa Accuracy SD Kappa SD

2 0.8940744 0.1866841 0.001269638 0.01934630

22 0.9059078 0.4866974 0.002218417 0.01546249

42 0.9053106 0.4904018 0.002470283 0.01106456

Accuracy was used to select the optimal model using the largest value.

The final value used for the model was mtry = 22.

The random forest has a slightly higher accuracy and a much higher kappa statistic, suggesting that it is a better model. The tuning picked an mtry (the number of variables available for splitting at each node)of 22. This model may be used for decisions since its statistics are pretty good.

**Wine Dataset**

I set out with the wine dataset to predict quality. First I look at summary statistics.

fixed.acidity volatile.acidity citric.acid residual.sugar chlorides

Min. : 4.60 Min. :0.1200 Min. :0.000 Min. : 0.900 Min. :0.01200

1st Qu.: 7.10 1st Qu.:0.3900 1st Qu.:0.090 1st Qu.: 1.900 1st Qu.:0.07000

Median : 7.90 Median :0.5200 Median :0.260 Median : 2.200 Median :0.07900

Mean : 8.32 Mean :0.5278 Mean :0.271 Mean : 2.539 Mean :0.08747

3rd Qu.: 9.20 3rd Qu.:0.6400 3rd Qu.:0.420 3rd Qu.: 2.600 3rd Qu.:0.09000

Max. :15.90 Max. :1.5800 Max. :1.000 Max. :15.500 Max. :0.61100

free.sulfur.dioxide total.sulfur.dioxide density pH sulphates

Min. : 1.00 Min. : 6.00 Min. :0.9901 Min. :2.740 Min. :0.3300

1st Qu.: 7.00 1st Qu.: 22.00 1st Qu.:0.9956 1st Qu.:3.210 1st Qu.:0.5500

Median :14.00 Median : 38.00 Median :0.9968 Median :3.310 Median :0.6200

Mean :15.87 Mean : 46.47 Mean :0.9967 Mean :3.311 Mean :0.6581

3rd Qu.:21.00 3rd Qu.: 62.00 3rd Qu.:0.9978 3rd Qu.:3.400 3rd Qu.:0.7300

Max. :72.00 Max. :289.00 Max. :1.0037 Max. :4.010 Max. :2.0000

alcohol quality

Min. : 8.40 Min. :3.000

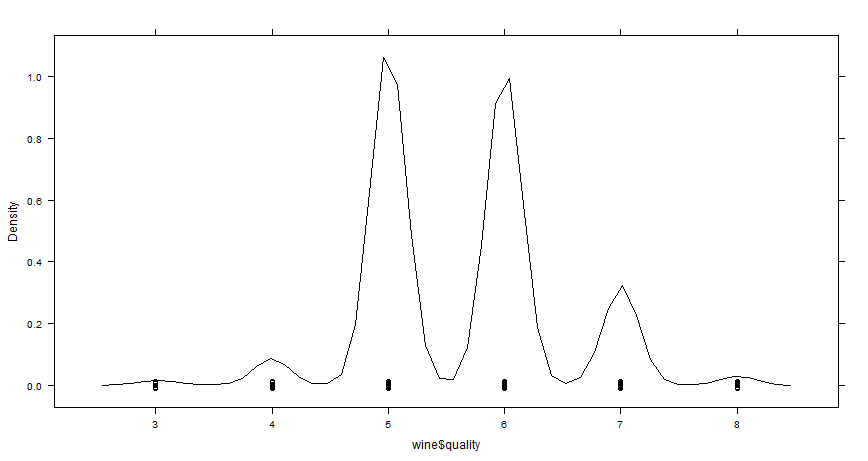
1st Qu.: 9.50 1st Qu.:5.000

Median :10.20 Median :6.000

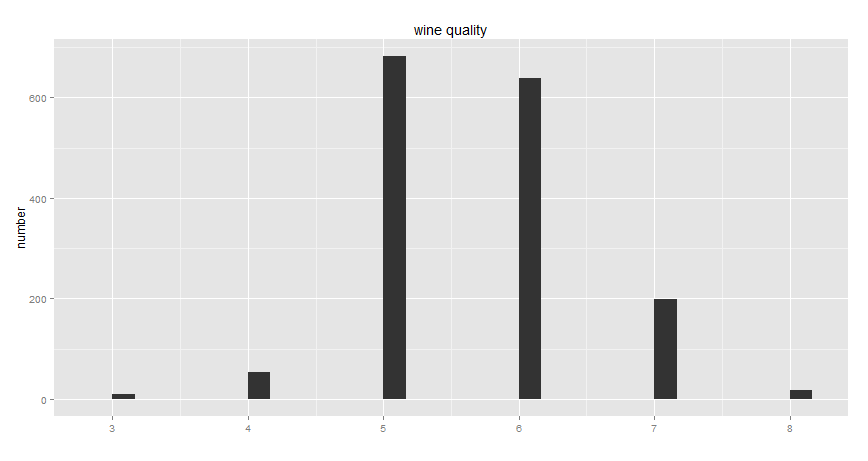
Mean :10.42 Mean :5.636

3rd Qu.:11.10 3rd Qu.:6.000

Max. :14.90 Max. :8.000

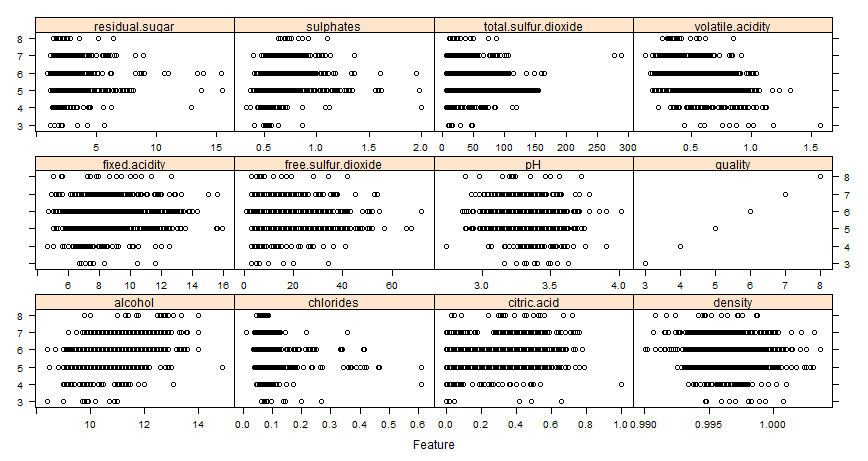


This distribution looks normal or bicameral.

The dataset looks approximately normal.

I move on to feature plots.

Feature Plots



It looks like we have some good predictors.

I begin with a random forest with 10-fold cross-validation, since that approach was successful last time.

Random Forest

1599 samples

11 predictor

No pre-processing

Resampling: Cross-Validated (10 fold)

Summary of sample sizes: 1439, 1440, 1440, 1438, 1438, 1440, ...

Resampling results across tuning parameters:

mtry RMSE Rsquared RMSE SD Rsquared SD

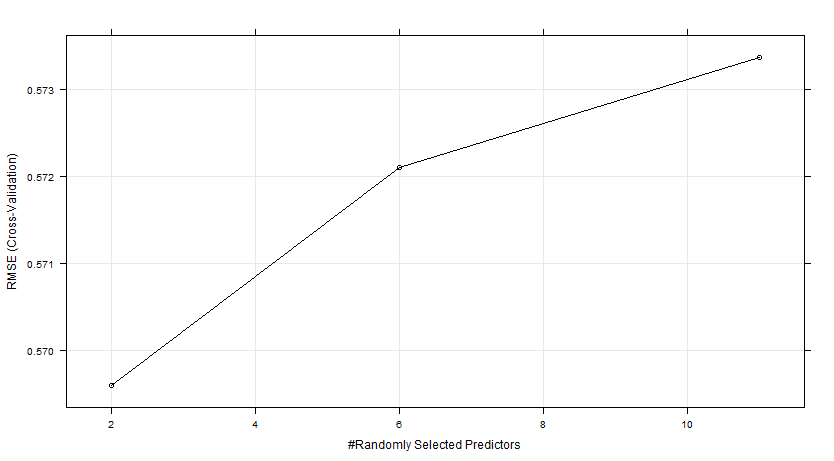
2 0.5696070 0.5148294 0.05896684 0.07238316

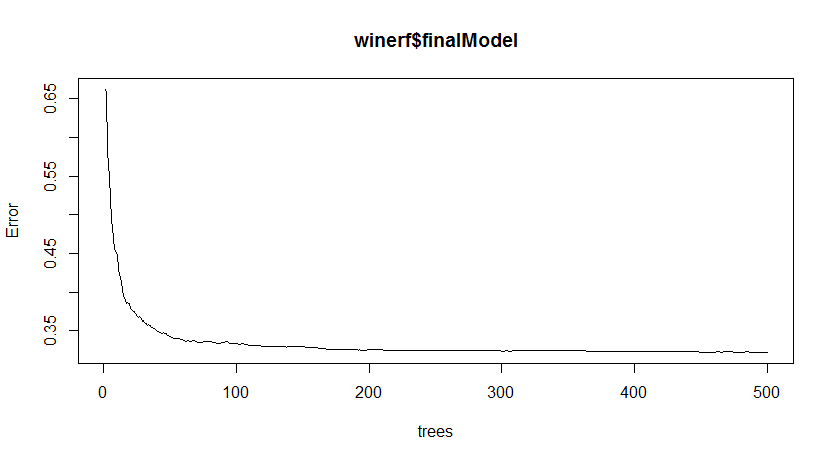
6 0.5721062 0.5014098 0.06033333 0.07514768

11 0.5733616 0.4981432 0.06138649 0.07557447

RMSE was used to select the optimal model using the smallest value.

The final value used for the model was mtry = 2.

Boosting Results



It looks like the number of trees converges around 200.

The model explains half of the variance and has an RMSE of .568. The tuning picked an mtry of 2. This model could be worse.

I move on to an SVM and a polynomial kernel whose degree will be selected by cross-validation

Support Vector Machines with Polynomial Kernel

1599 samples

11 predictor

No pre-processing

Resampling: Cross-Validated (10 fold)

Summary of sample sizes: 1440, 1439, 1440, 1438, 1440, 1438, ...

Resampling results across tuning parameters:

degree scale C RMSE Rsquared RMSE SD Rsquared SD

1 0.001 0.25 0.7124468 0.3327484 0.02214550 0.08188166

1 0.001 0.50 0.6786760 0.3437628 0.02595026 0.07992068

1 0.001 1.00 0.6641469 0.3502716 0.02784971 0.07755284

1 0.010 0.25 0.6564121 0.3508737 0.03006923 0.07636226

1 0.010 0.50 0.6553889 0.3500887 0.03252780 0.07537213

1 0.010 1.00 0.6549953 0.3507889 0.03445096 0.07612394

1 0.100 0.25 0.6559494 0.3502643 0.03570704 0.07678044

1 0.100 0.50 0.6563543 0.3496435 0.03625268 0.07699504

1 0.100 1.00 0.6565561 0.3493263 0.03639053 0.07678210

2 0.001 0.25 0.6784671 0.3441981 0.02600730 0.08021828

2 0.001 0.50 0.6640172 0.3506609 0.02785867 0.07781696

2 0.001 1.00 0.6571875 0.3516562 0.02945188 0.07728335

2 0.010 0.25 0.6487359 0.3627638 0.03455764 0.08181529

2 0.010 0.50 0.6459653 0.3668144 0.03634497 0.08189994

2 0.010 1.00 0.6456797 0.3677551 0.03839881 0.08297111

2 0.100 0.25 0.6512254 0.3629311 0.04165640 0.08246126

2 0.100 0.50 0.6533389 0.3603779 0.04168852 0.08076031

2 0.100 1.00 0.6563912 0.3562942 0.04188735 0.07925027

3 0.001 0.25 0.6685292 0.3498039 0.02697502 0.07881679

3 0.001 0.50 0.6583174 0.3522836 0.02865341 0.07770404

3 0.001 1.00 0.6548845 0.3526184 0.03104489 0.07721279

3 0.010 0.25 0.6452311 0.3682051 0.03622882 0.08175890

3 0.010 0.50 0.6448310 0.3692333 0.03835820 0.08275376

3 0.010 1.00 0.6452414 0.3696660 0.04054294 0.08454448

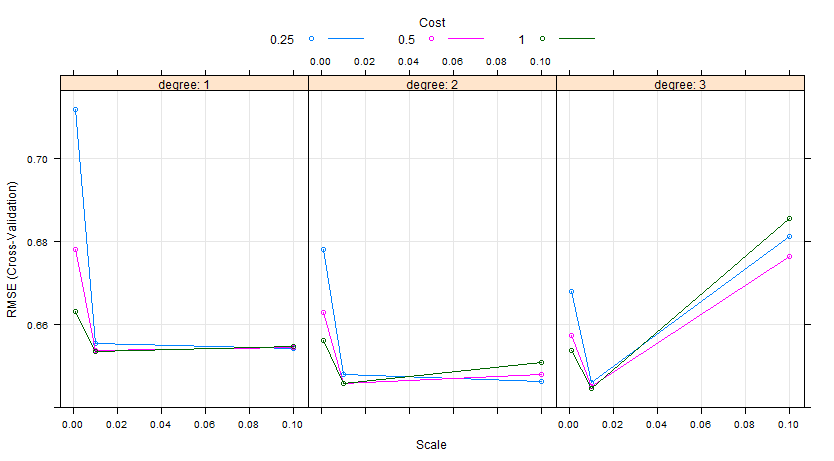
3 0.100 0.25 0.6894090 0.3282343 0.07871782 0.07966856

3 0.100 0.50 0.6884765 0.3354071 0.08488555 0.07534313

3 0.100 1.00 0.6963403 0.3342062 0.09986420 0.07999152

RMSE was used to select the optimal model using the smallest value.

The final values used for the model were degree = 3, scale = 0.01 and C = 0.5.

Boosting ResultsThe SVM has a greater RMSE and lower R-squared, so it is a worse model than the random forest. The tuning picked a degree 3, scale .01, and c of 0.5. I would use the random forest for decision making on the wine dataset over the SVM.