Binary classification problem

Jonathan P. Latner, PhD

June 19, 2022

Overview

Goal

- Explore both data sets, note down your key observations along with a kind of summary.
- Build a classifier – a prediction model based only on the training data, with the goal of achieving the best performance possible on the validation data.
- Visualize results and the work on this classification task.

Figure 1: Factor variables

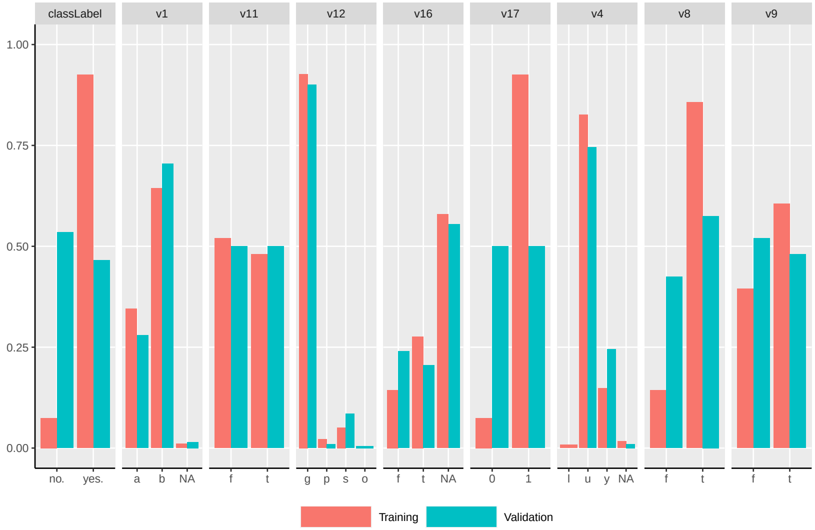


Figure 2: Factor variables (correlation)

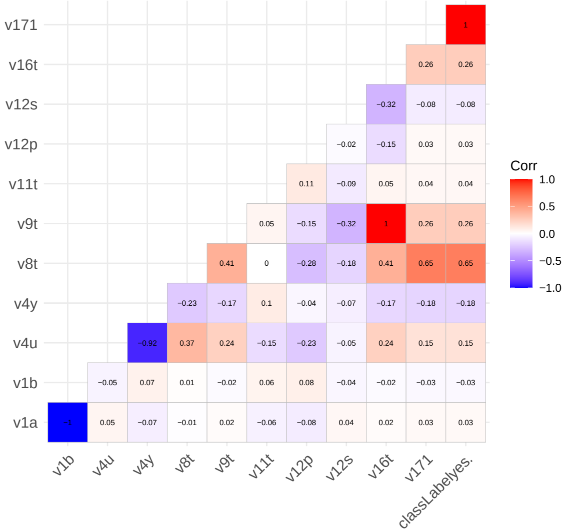


Figure 3: Numerical variables

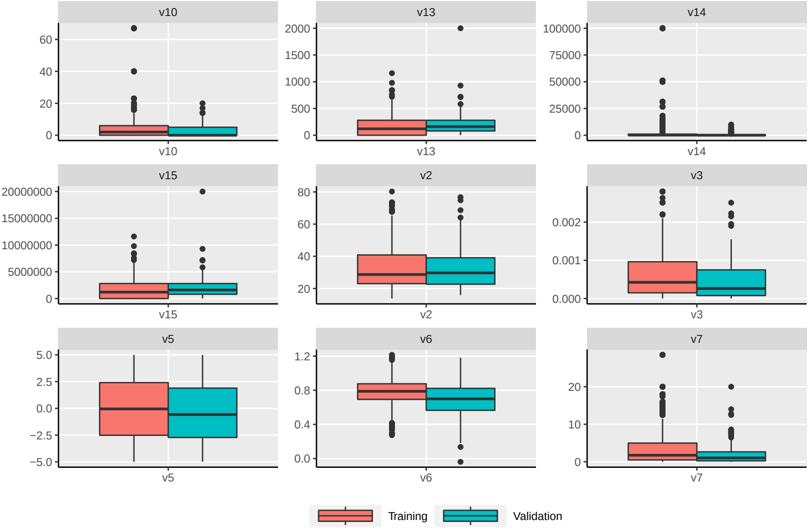


Figure 4: Numerical variables (correlation)

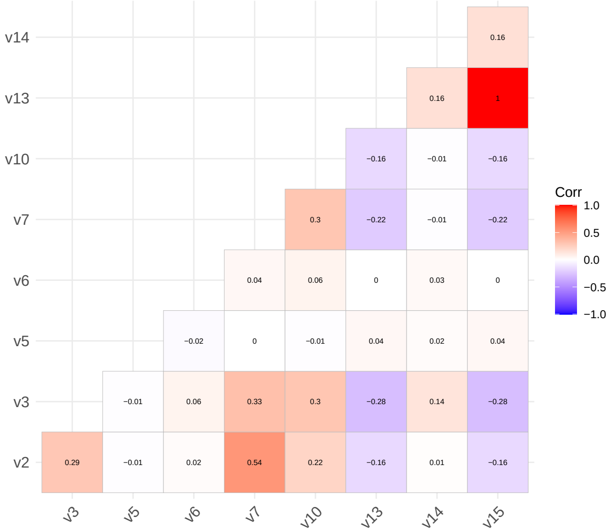
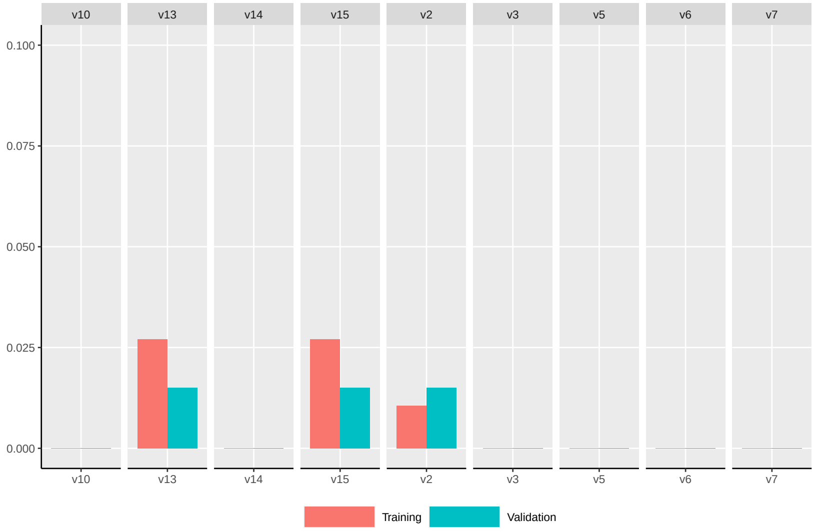


Figure 5: Numerical variables (missing)



Initial lessons

- Classlabel (dv)
  - v17 = classLabel: keep classLabel, drop v17
  - Difference in distribution between training and validation data
  - Questionable power to predict 0/no, given low count in training data
- v3 has mean, median, mode, and sd of 0, drop v3
- v12 training data has no "o"
- v4 validation data has no "l"
- v14 and especially v15 have a very long tail
- v13 and v15 perfectly correlated, drop v15
- Variable v16
  - v9 and v16 are perfectly correlated (non missings)
  - v16 has lots of missing observations (but missing in both traning and validation data)
  - or is v9, v16 without any missing?

Steps

- Compare different models with all IVs
  - a) GLM
  - b) Decision tree
  - c) Random forest
  - d) Naive Bayes
- Examine model fit, summary statistics, etc.

Confusion matrix

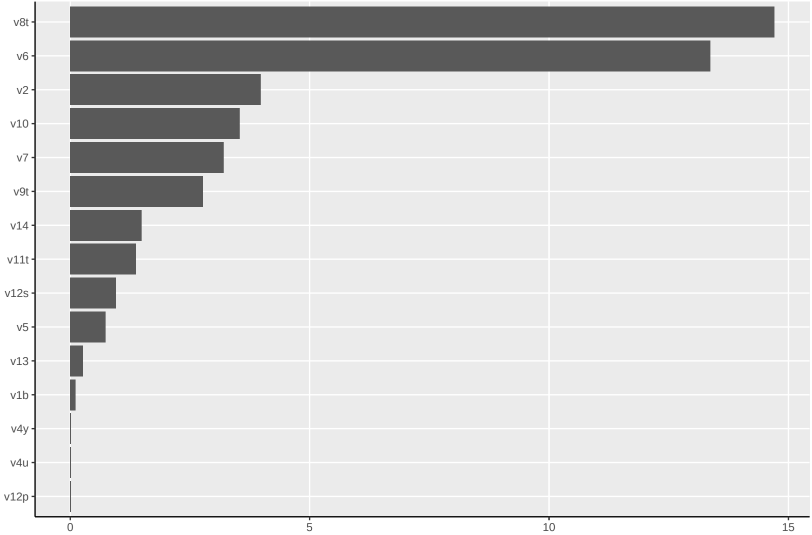
		Model 1a: GLM	
reality	predicted	Freq	Pct
0	0	49	0.91
1	0	5	0.09
0	1	50	0.37
1	1	86	0.63
Accuracy		0.711	
Duration (secs)		2.06	

Table 1: Parameter estimates from logistic regression model 1a

	Model 1
(Intercept)	7.85 (838.37)
v1b	0.02 (0.23)
v2	-0.04 (0.01)***
v4u	-15.33 (838.37)
v4y	-15.68 (838.37)
v5	0.03 (0.03)
v6	11.99 (0.90)***
v7	0.16 (0.05)**
v8t	3.58 (0.24)***
v9t	-0.91 (0.33)**
v10	0.28 (0.08)***
v11t	0.29 (0.21)
v12p	-14.86 (1214.23)
v12s	-0.33 (0.34)
v13	0.00 (0.00)
v14	0.00 (0.00)
Log Likelihood	-359.09
Num. obs.	3523

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Figure 6: Variable importance



Lessons from model GLM

- Decent model fit
  - Good at predicting 0
  - Bad at predicting 1
- Drop variables?
  - Categorical - v1, v4, combine v12s/v12p
  - Continuous - v13, v14

Add decision tree

		Model 1a: GLM		Model 1b: Tree	
reality	predicted	Freq	Pct	Freq	Pct
0	0	49	0.91	51	0.96
1	0	5	0.09	2	0.04
0	1	50	0.37	48	0.35
1	1	86	0.63	89	0.65
Accuracy		0.711		0.737	
Duration (secs)		2.06		2.05	

Lessons from model 2

- Slightly better fit
- Slightly better at predicting 0 (model 1 already good)
- Slightly better, but still bad at predicting 1

Add random forest

		Model 1a: GLM		Model 1b: Tree		Model 1c: RF	
reality	predicted	Freq	Pct	Freq	Pct	Freq	Pct
0	0	49	0.91	51	0.96	72	0.95
1	0	5	0.09	2	0.04	4	0.05
0	1	50	0.37	48	0.35	27	0.24
1	1	86	0.63	89	0.65	87	0.76
Accuracy		0.711		0.737		0.837	
Duration (secs)		2.06		2.05		65.47	

Add naive bayes

		Model 1a: GLM		Model 1b: Tree		Model 1c: RF		Model 1d: NB	
reality	predicted	Freq	Pct	Freq	Pct	Freq	Pct	Freq	Pct
0	0	49	0.91	51	0.96	72	0.95	41	0.98
1	0	5	0.09	2	0.04	4	0.05	1	0.02
0	1	50	0.37	48	0.35	27	0.24	58	0.39
1	1	86	0.63	89	0.65	87	0.76	90	0.61
Accuracy		0.711		0.737		0.837		0.689	
Duration (secs)		2.06		2.05		65.47		18.82	

Summary

- Random forest regression offers best fit
- Also much, much slower (time/energy costs money)
- Next steps: improve the model

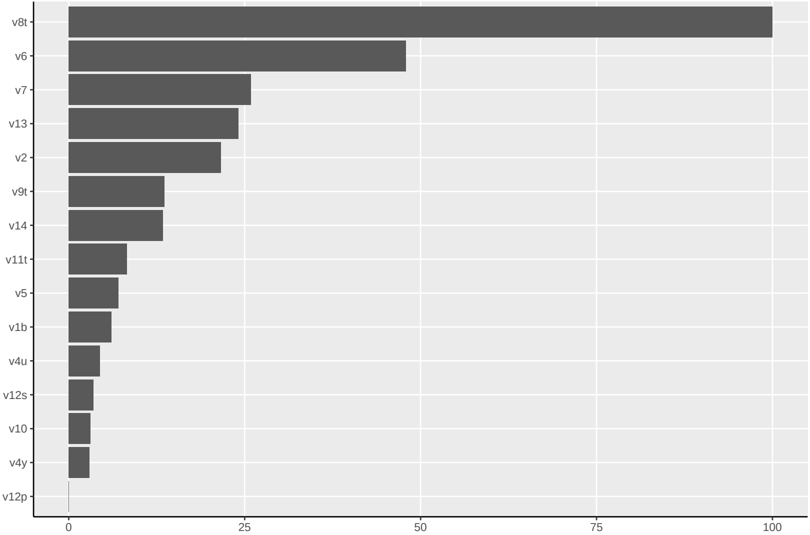
Steps

- Begin with random forest baseline (all IVs, same as model 1c)
- Examine model fit, summary statistics, etc.
- Make adjustments
- Rerun model
- Repeat as necessary

Confusion matrix

		Model 2a: Base	
reality	predicted	Freq	Pct
0	0	71	0.95
1	0	4	0.05
0	1	28	0.24
1	1	87	0.76
Accuracy		0.832	
Duration (secs)		93.49	

Figure 7: Variable importance



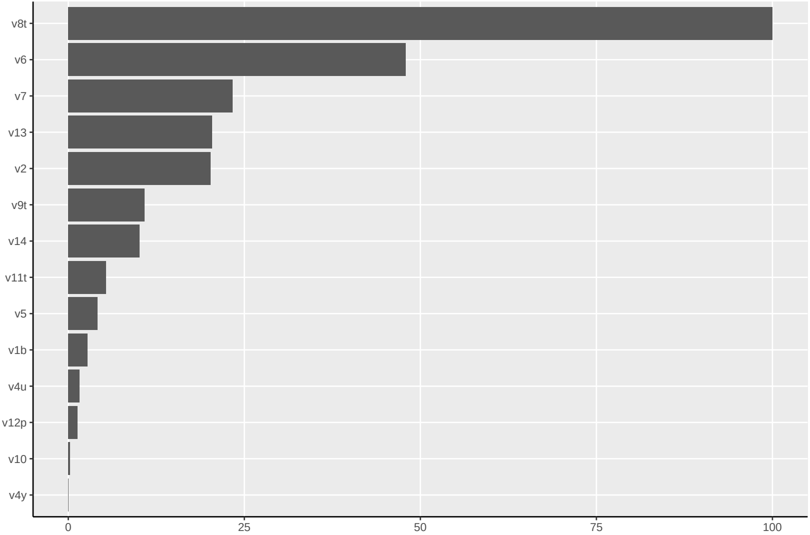
Lessons

- Combine v12s/v12p

Confusion matrix

		Model 2a: Base		Model 2b: v12	
reality	predicted	Freq	Pct	Freq	Pct
0	0	71	0.95	69	0.96
1	0	4	0.05	3	0.04
0	1	28	0.24	30	0.25
1	1	87	0.76	88	0.75
Accuracy		0.832		0.826	
Duration (secs)		93.49		82.03	

Figure 8: Variable importance



Lessons

- Slightly worse fit, but faster
- Combine v4y/v4u
- Drop v10

Confusion matrix

		Model 2a: Base		Model 2b: v12		Model 2c: v4/v10	
reality	predicted	Freq	Pct	Freq	Pct	Freq	Pct
0	0	71	0.95	69	0.96	67	0.94
1	0	4	0.05	3	0.04	4	0.06
0	1	28	0.24	30	0.25	32	0.27
1	1	87	0.76	88	0.75	87	0.73
Accuracy		0.832		0.826		0.811	
Duration (secs)		93.49		82.03		60.87	

Lessons

- Slightly worse fit, but faster

- Random forest is preferable model
  - Good at predicting 0 (4% false negative)
  - Okay at predicting 1 (25% false positive)
- Model includes following variable modifications:
  - v17 = classLabel (drop v17)
  - assume that v9 = v16, but without missing (drop v16)
  - v15 = v13, perfectly correlated (drop v15)
  - Drop v3 due to no variation (mean, median, mode, and sd = 0)
  - v12s = v12p (combine values)
  - Drop v12o (in validation, but not training data)
- Next steps: model improvement
  - Focus on what predicts 1
  - Interacting variables
    - v8 and v9?
  - Binning continuous variables
    - v10, v13, v14 all have lots of 0's

Thank you