ASC Conference

Ort House, London, 15 November 2018

Validating Black Box Neural Networks

Dale Chant, Michael Potter, Red Centre Software

Contents:

- Why need to validate?
- Independent Benchmarks
- Predicting Likely to Recommend (NPS) from expatiate Sentiment
  - Generic verbatim issues
  - A verifiable algorithm: syuzhet* Benchmark
  - Machine Learning at different resolutions (binary to full scale)
  - Verify ML predictions by
    - Case plot against training set
    - Case plot against syuzhet score
    - Correlation to training data

*syuzhet: an R package for text analysis. In Russian formalism, *syuzhe*t is the narrative structure of a novel/drama/poem.

Why need to Validate?

- Machine learning (most commonly neural nets) is practically opaque to a debug walk-through
- Reruns against the same training data can yield quite different models,

because

- Layers are initialised by random weights and biases,

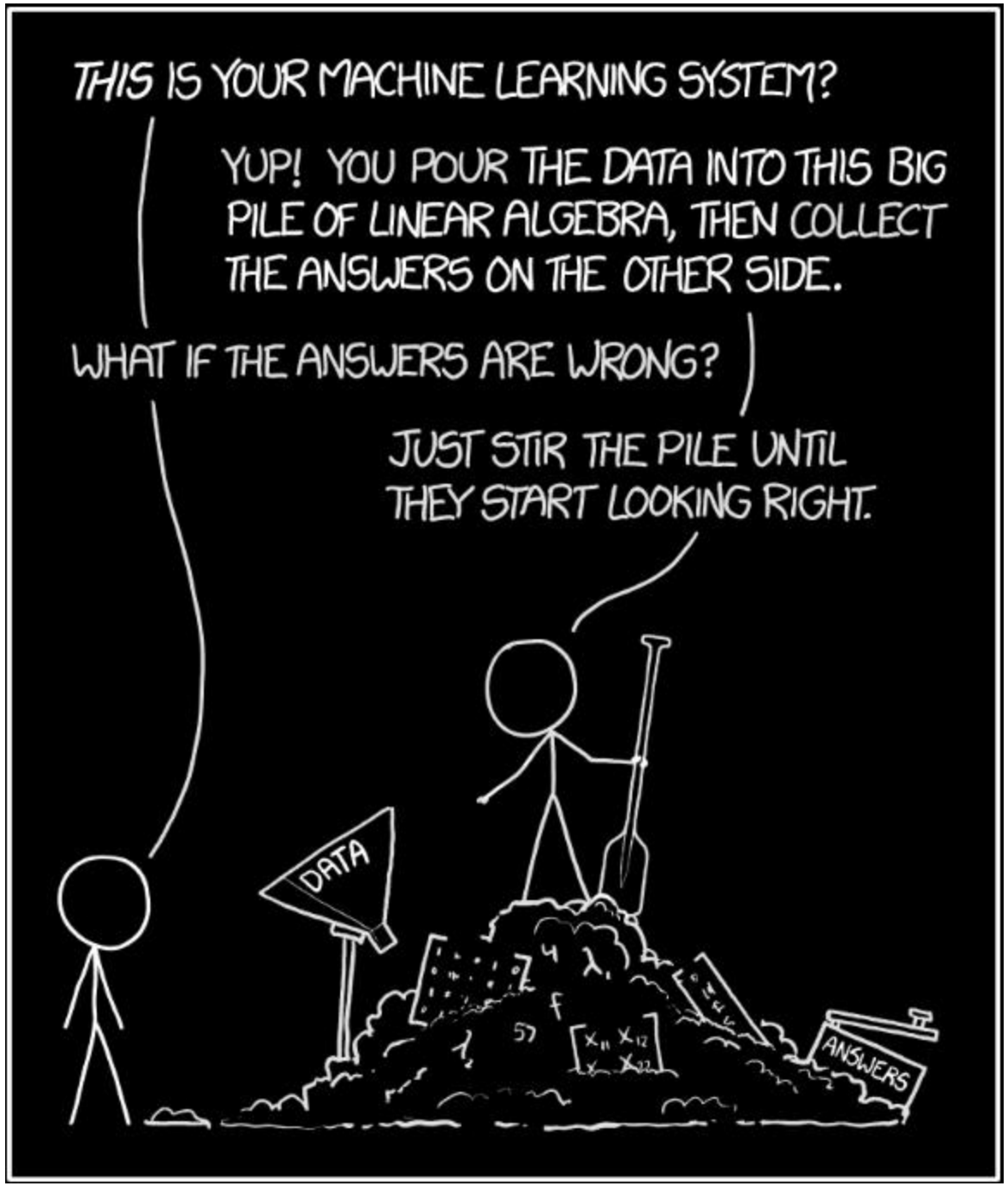
so

- No guarantee the cost function has not found misleading local minima

Therefore, one cannot explain why or how a particular prediction was correct/incorrect

NB! We make no claims to ML/NN expertise. This is an empirical account only of our experiences in trying to get to something sensible and useful.

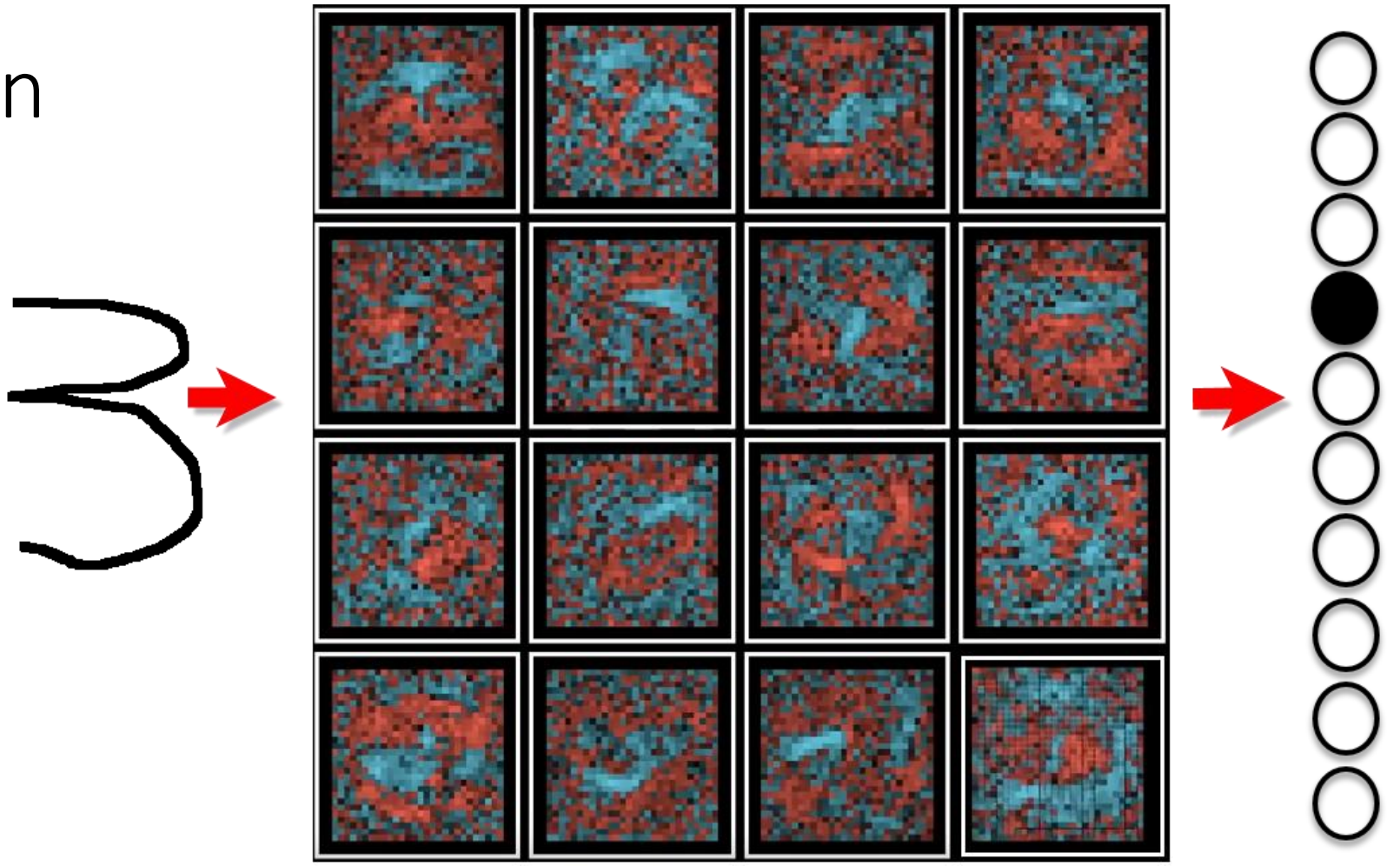
Why Need to Validate:



Not reassuring for a client

https://xkcd.com/1838/

Hand-written Digits



Pixelated input **Pixel representation of hidden layers** **Fires 4th (0-based=3)**

3blue1brown: https://youtu.be/IHZwWFHWa-w

To Validate: Contrive Independent Benchmarks

- For hand-written digits:
  - Is the output better than a random guess (one in ten chance of being correct)
  - A “1” has far fewer black pixels (on white background) than “8”. A “7” about half as many. Measuring ‘darkness’ gives about 50% correct

- For Brand Coding:
  - Compare the correct rate to that obtained from Approximate String Matching algorithms such as Damerau-Levenshtein
  - Correlation of trained vs predicted on same cases

- For text sentiment:
  - Compare the correct rate to that obtained from summing pre-defined word weights
  - Correlation of trained vs predicted on same cases

To Validate: Test for Consistent Model Behaviour

- Test the model on its own training data
  - Having already seen the answers, how well does it remember?
- Compare multiple runs
  - Disparities measure instability
- Does the ratio of correct:total degrade naturally as resolution increases?
  - One would expect Binary sentiment (predict just positive or negative) to be a great deal more accurate than predicting a value from 1 to 5, and that in turn more accurate than 0 to 10

Predicting *Likely to Recommend* (NPS) from Sentiment

- Q4a: *On a scale of 0 to 10, how likely are you to recommend <brand> to friend or relative? [8]*
- Q4b: *Why did you give that rating? ["Good service, reasonable price"]*

Rating + expatiate questions provide a self-tagged training set. This respondent equates the sentiment of "*Good service, reasonable price*" to be = *8*

- **Given just the text *Good service, reasonable price*, how reliably can the sentiment score = *8* be predicted?**

Generic Verbatim Issues

- Truncated by field:
  *Low annual premiums; option of no excess; ease of using hospital cover, i.e. bills sent directly **fro***

- Slang and abbreviations:
  *I **wanna** leave and change **asap***

- Formatting:
  Happywiththeirsystem

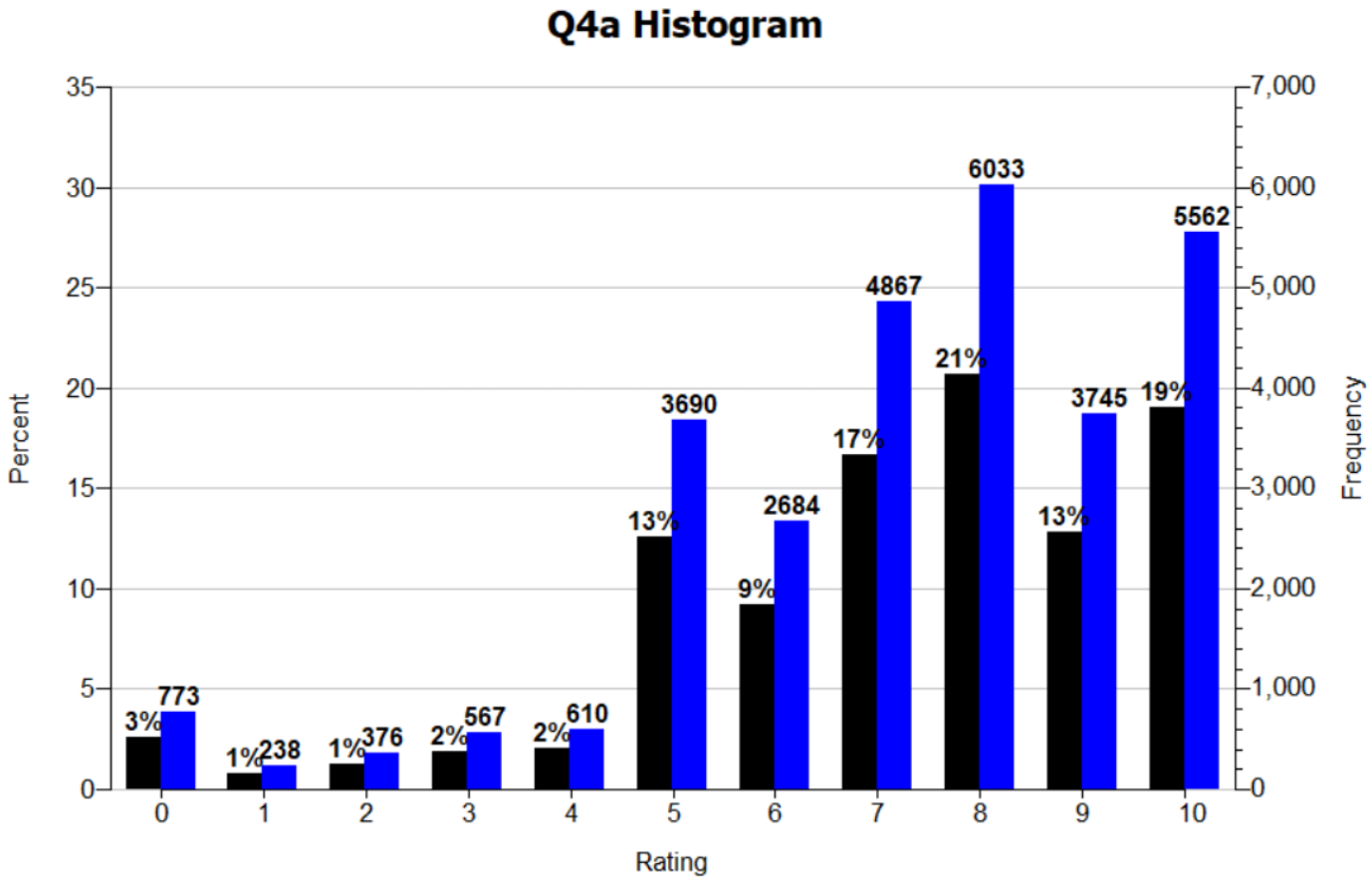
- Nonsense:
  Daddy is good at it

- Puzzling:
  confiability

- Skimmers:
  434774244347742443477

- Spelling:
  physcologist

Q4a Distribution



- Over 90% of the 29,000 respondents gave a rating >= 5
- Ratings 7, 8 and 10 most popular
- Negative/positive apex between 6 and 7 (by manual review)

The Independent Benchmark: Syuzhet Scores

syuzhet: an R package for text analysis. In Russian formalism, *syuzhe*t is the narrative structure of a novel/drama/poem.

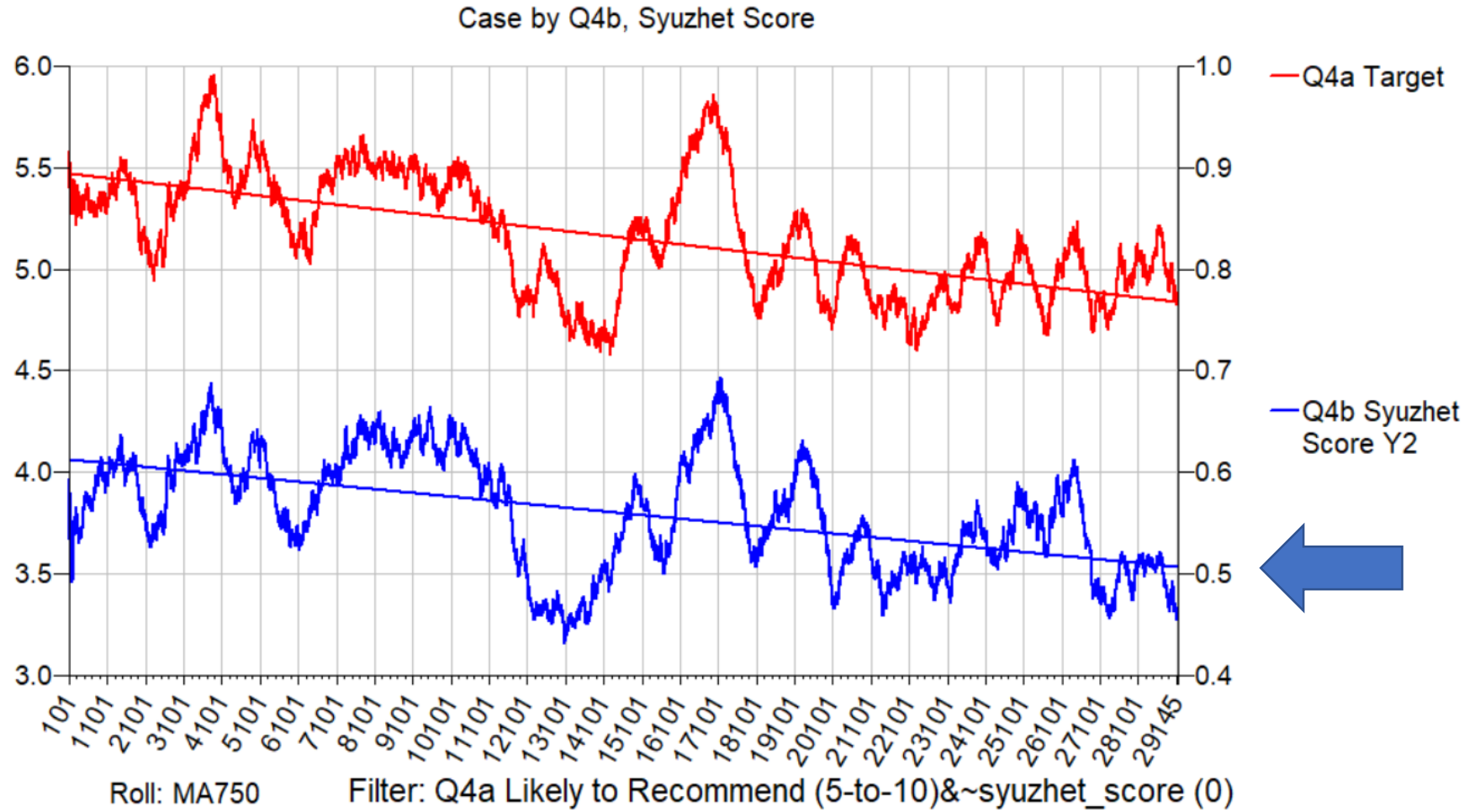
- How it works:
  if love=0.75, beauty=0.5, good=0.5 then *Love of beauty is good* scores 0.75+0.5+0.5 = 1.75

- Most negative response at score of -4:
  …lack of details concerning claimable areas…non-claimable costs…feels cheap and nasty to say a health service is claimable, only to discover it's not…disappointing and covert

- Most positive response at score of 8.4:
  … The yearly cost was half … the benefits were immensely better…the best value available today… not for profit fund… accessible for helpful information...

- Syuzhet issues:
  - But many zeroes because
    - Word misspelled so weight lookup fails
    - Words not in dictionary
    - Opposites can cancel
  - Long responses can accumulate greater absolute score

The Syuzhet Benchmark

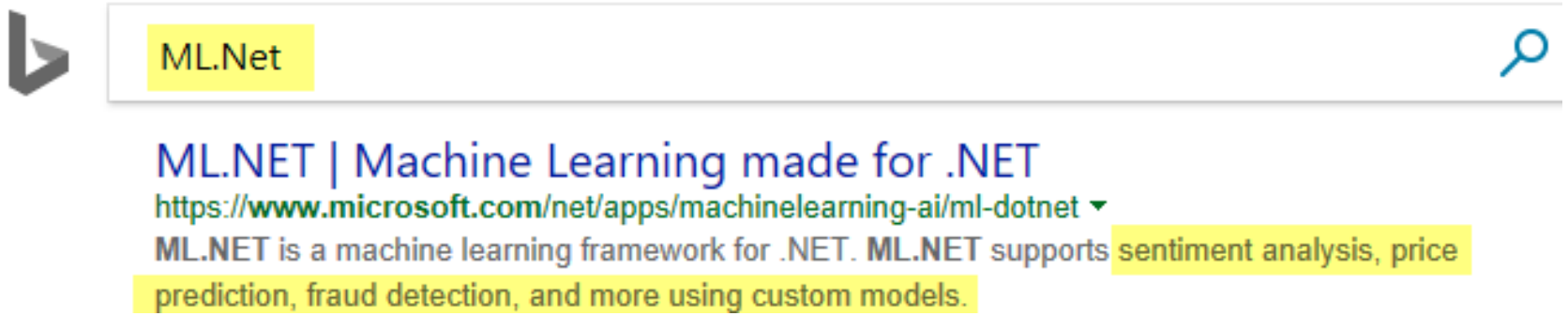


Can ML do better than this?

- Red is target, being the respondents' own ratings
- Blue is syuzhet on spell/grammar-checked (un-spell/grammar-checked is almost as good)
- De-noised by filter and roll

Using Microsoft's ML.Net

- Free AI/ML libraries
- Caller code (get inputs, apply models, write outputs) in VB.Net, compiled in Visual Studio 2017 Community (free if <$USmillion) – search for *ML.Net*



- An agency-friendly environment – MS VB on Windows desktop rather than C/C++ or Python under Linux
- We experimented with and adapted the supplied examples

The Model Inputs

- For each respondent, assemble self-rating + verbatim as

0 to 10 scale

```
8       "LOWER THEIR PREMIUMS - MAYBE MORE PEOPLE WOULD
7       "PROVIDE CHEAPER PREMIUMS"      1
8       "UPDATE THEIR DENTAL COVERAGE"  1
10      "EASY TO UNDERSTAND. CONSULTANTS ARE REALLY FRIE
9       "I HAVE NEVER EVER IN 12 YRS HAD A PROBLEM"
6       "Unsure"        1
10      "THEY PAY OUT AND LOOK AFTER US"        1
8       "COVER 100%"    1
```

0 to 1 (binary scale) as net of 0 to 6 = 0 = negative, and 7 to 10 = 1 = positive

```
1       "Unsure"
1       "THEY PAY OUT AND LOOK AFTER US"
1       "COVER 100%"
1       "THEIR COST"
1       "MAKE COSTS CHEAPER"
0       "POOR DENTAL COVERAGE"
1       "THEY ARE EASY TO DEAL WITH .THEY HAVE BRANCHES EVERYWHERE, THEY ARE BACKED BY THE GOVT
0       "I THINK THAT HEALTH INSURANCE IS AN EVIL NECESSITY SO I GUESS I WOULDN'T RECOMMEND IT
0       "Not sure"
0       "EVERYONE HAS DIFFERENT NEEDS, THEY SHOULD LOOK INTO THE HEALTH INSURANCE THAT SUITS TH
1       "ALWAYS HELPFUL AND POLITE WHEN ASKED QUESTIONS  REASONABLY PRICED"
```

Run the Model

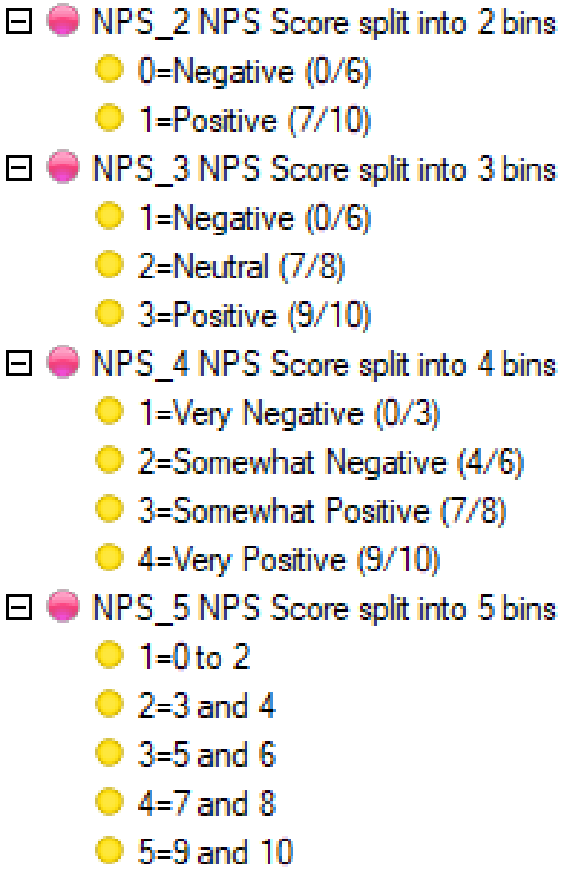
- Loop five times for comparisons

```
For runID = 1 To 5
      model = TrainSentimentMultiClass("Q4a_1", InputTextVar, TrainCaseFilter)
      EvaluateSentimentMultiClass(model, … InputTextVar, "Q4a_1", EvalCaseFilter, runID)
Next


Public Function TrainSentimentMultiClass(…)
      …
     Dim sdcac As New Trainers.StochasticDualCoordinateAscentClassifier
      …
     model = pipeline.Train(Of ClassificationData, ClassPredictionMultiClass)

      Return model
End Function


Public Function EvaluateSentimentMultiClass(model…)
     …
    pred = model.Predict(probe)
     …
End Function
```

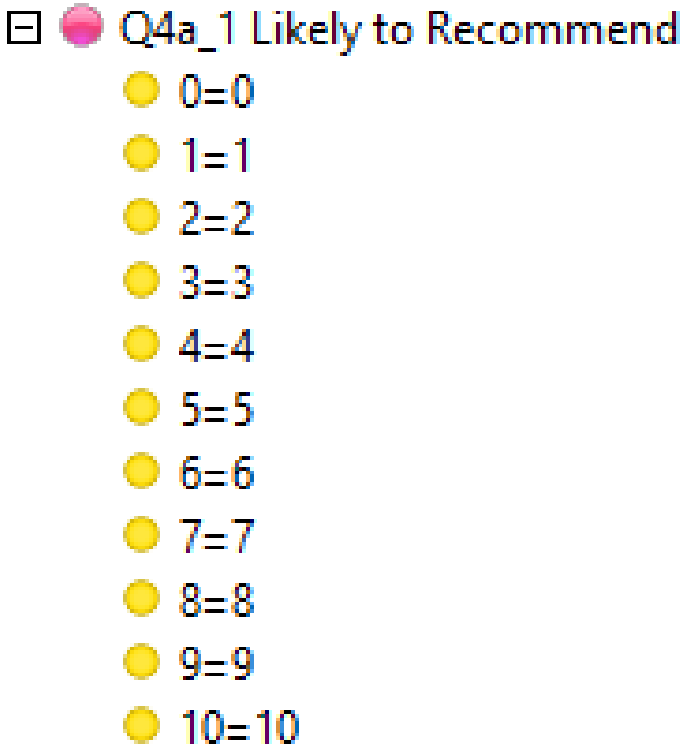
The BLACK BOX

Compare Outputs

- Models were executed five times at each resolution (2, 3, 4, 5, 11)



Constructed targets as nets of the source Q4a



Res11 target is Q4a itself

At Lowest Resolution (binary)

Correctly predicted : 96.30%

Column Percents	Predicted Res2 Run1	
	Negative (0/6)	Positive (7/10)
Cases	9,060	20,085
Q4a Likely to Recommend 0	8%	0%
1	3%	0%
2	4%	0%
3	6%	0%
4	6%	0%
5	39%	1%
6	28%	1%
7	2%	23%
8	3%	29%
9	1%	18%
10	1%	27%
Net 0 to 6	93%	2%
Net 7 to 10	7%	98%

Correctly predicted : 96.31%

Column Percents	Predicted Res2 Run2	
	Negative (0/6)	Positive (7/10)
Cases	9,059	20,086
Q4a Likely to Recommend 0	8%	0%
1	3%	0%
2	4%	0%
3	6%	0%
4	6%	0%
5	39%	1%
6	28%	1%
7	2%	23%
8	3%	29%
9	1%	18%
10	1%	27%
Net 0 to 6	93%	2%
Net 7 to 10	7%	98%

Correctly predicted : 96.30%

Column Percents	Predicted Res2 Run3	
	Negative (0/6)	Positive (7/10)
Cases	9,061	20,084
Q4a Likely to Recommend 0	8%	0%
1	3%	0%
2	4%	0%
3	6%	0%
4	6%	0%
5	39%	1%
6	28%	1%
7	2%	23%
8	3%	29%
9	1%	18%
10	1%	27%
Net 0 to 6	93%	2%
Net 7 to 10	7%	98%

Correctly predicted : 96.31%

Column Percents	Predicted Res2 Run4	
	Negative (0/6)	Positive (7/10)
Cases	9,062	20,083
Q4a Likely to Recommend 0	8%	0%
1	3%	0%
2	4%	0%
3	6%	0%
4	6%	0%
5	39%	1%
6	28%	1%
7	2%	23%
8	3%	29%
9	1%	18%
10	1%	27%
Net 0 to 6	93%	2%
Net 7 to 10	7%	98%

Correctly predicted : 96.31%

Column Percents	Predicted Res2 Run5	
	Negative (0/6)	Positive (7/10)
Cases	9,061	20,084
Q4a Likely to Recommend 0	8%	0%
1	3%	0%
2	4%	0%
3	6%	0%
4	6%	0%
5	39%	1%
6	28%	1%
7	2%	23%
8	3%	29%
9	1%	18%
10	1%	27%
Net 0 to 6	93%	2%
Net 7 to 10	7%	98%

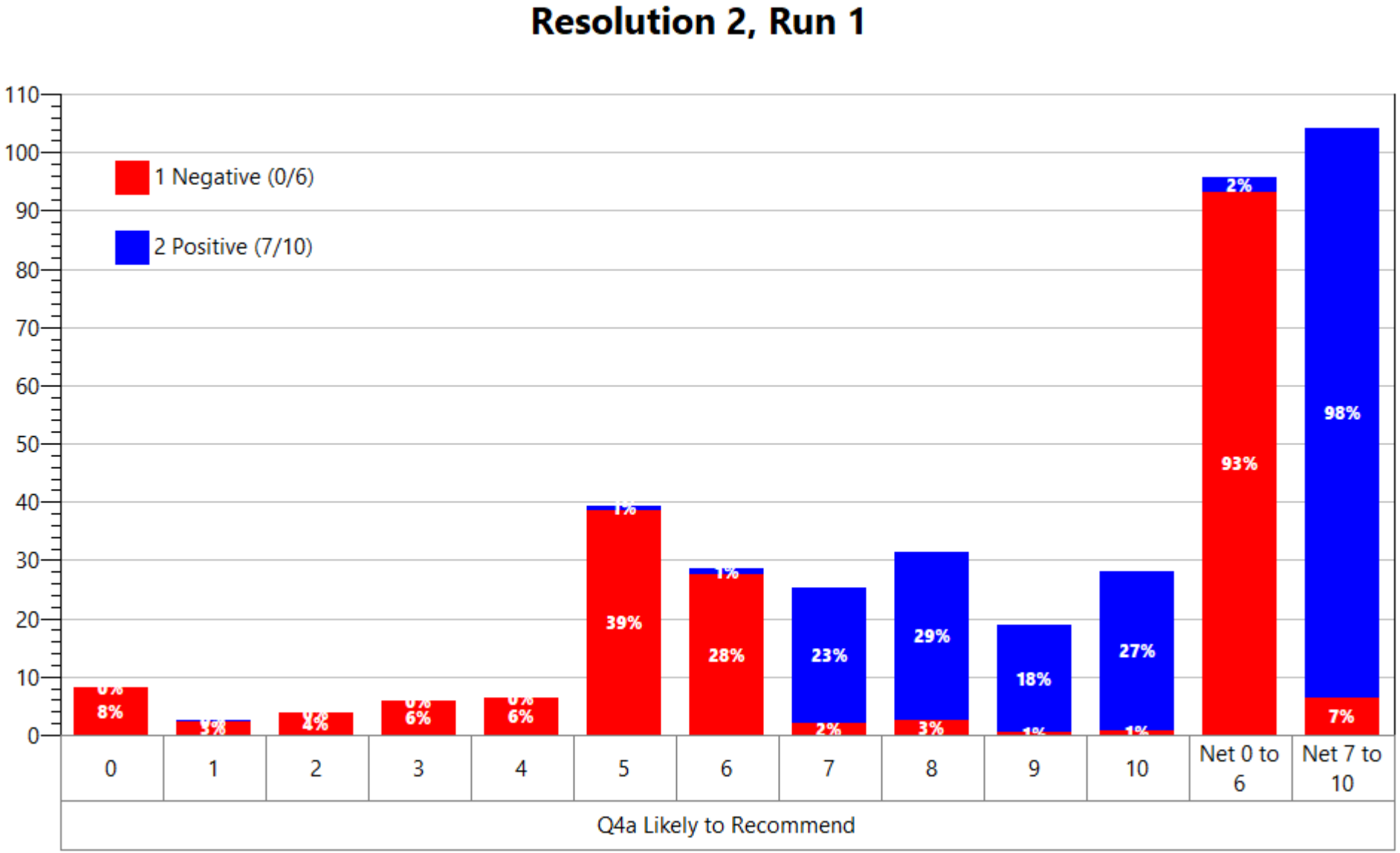
Correctly predicted : 96.30% =100-100*(600+478)/(9060+20085)

Frequencies Column Percents	Predicted Res2 Run1	
	Negative (0/6)	Positive (7/10)
Q4a Likely to Recom Cases	9,060	20,085
Net 0 to 6	8,460	478
	93%	2%
Net 7 to 10	600	19,607
	7%	98%

	Min	Max	Diff
Negative	9059	9062	**3**
Positive	20083	20086	**3**

- **Very stable (identical %s at 0DP)**
- **But binary is the widest possible target**

As Chart:


Resolution 2, Run 1
1 Negative (0/6)
2 Positive (7/10)
110
100
90
80
70
60
50
40
30
20
10
0
8%
4%
6%
6%
39%
1%
28%
1%
23%
2%
29%
3%
18%
27%
1%
2%
93%
98%
7%
0
1
2
3
4
5
6
7
8
9
10
Net 0 to 6
Net 7 to 10
Q4a Likely to Recommend

At Quad Resolution

Correctly predicted : 74.11%

Column Percents		Predicted Res4 Run1			
		Very Neg (0/3)	Some Neg (4/6)	Some Pos (7/8)	Very Pos (9/10)
Q4a Likely to Rec	Cases	346	7,640	11,754	9,405
	Net 0 to 3	61%	14%	4%	2%
	Net 4 to 6	29%	66%	11%	5%
	Net 7 to 8	8%	14%	74%	12%
	Net 9 to 10	3%	6%	10%	81%

Correctly predicted : 74.24%

Column Percents		Predicted Res4 Run2			
		Very Neg (0/3)	Some Neg (4/6)	Some Pos (7/8)	Very Pos (9/10)
Q4a Likely to Rec	Cases	909	7,010	11,294	9,932
	Net 0 to 3	47%	13%	3%	2%
	Net 4 to 6	36%	68%	11%	6%
	Net 7 to 8	11%	13%	76%	12%
	Net 9 to 10	6%	6%	9%	79%

Correctly predicted : 74.00%

Column Percents		Predicted Res4 Run3			
		Very Neg (0/3)	Some Neg (4/6)	Some Pos (7/8)	Very Pos (9/10)
Q4a Likely to Rec	Cases	93	9,064	11,076	8,912
	Net 0 to 3	71%	15%	3%	2%
	Net 4 to 6	18%	62%	9%	4%
	Net 7 to 8	5%	15%	77%	11%
	Net 9 to 10	5%	8%	11%	83%

Correctly predicted : 74.42%

Column Percents		Predicted Res4 Run4			
		Very Neg (0/3)	Some Neg (4/6)	Some Pos (7/8)	Very Pos (9/10)
Q4a Likely to Rec	Cases	147	8,352	10,944	9,702
	Net 0 to 3	63%	15%	3%	2%
	Net 4 to 6	27%	64%	10%	5%
	Net 7 to 8	5%	14%	78%	12%
	Net 9 to 10	4%	7%	9%	80%

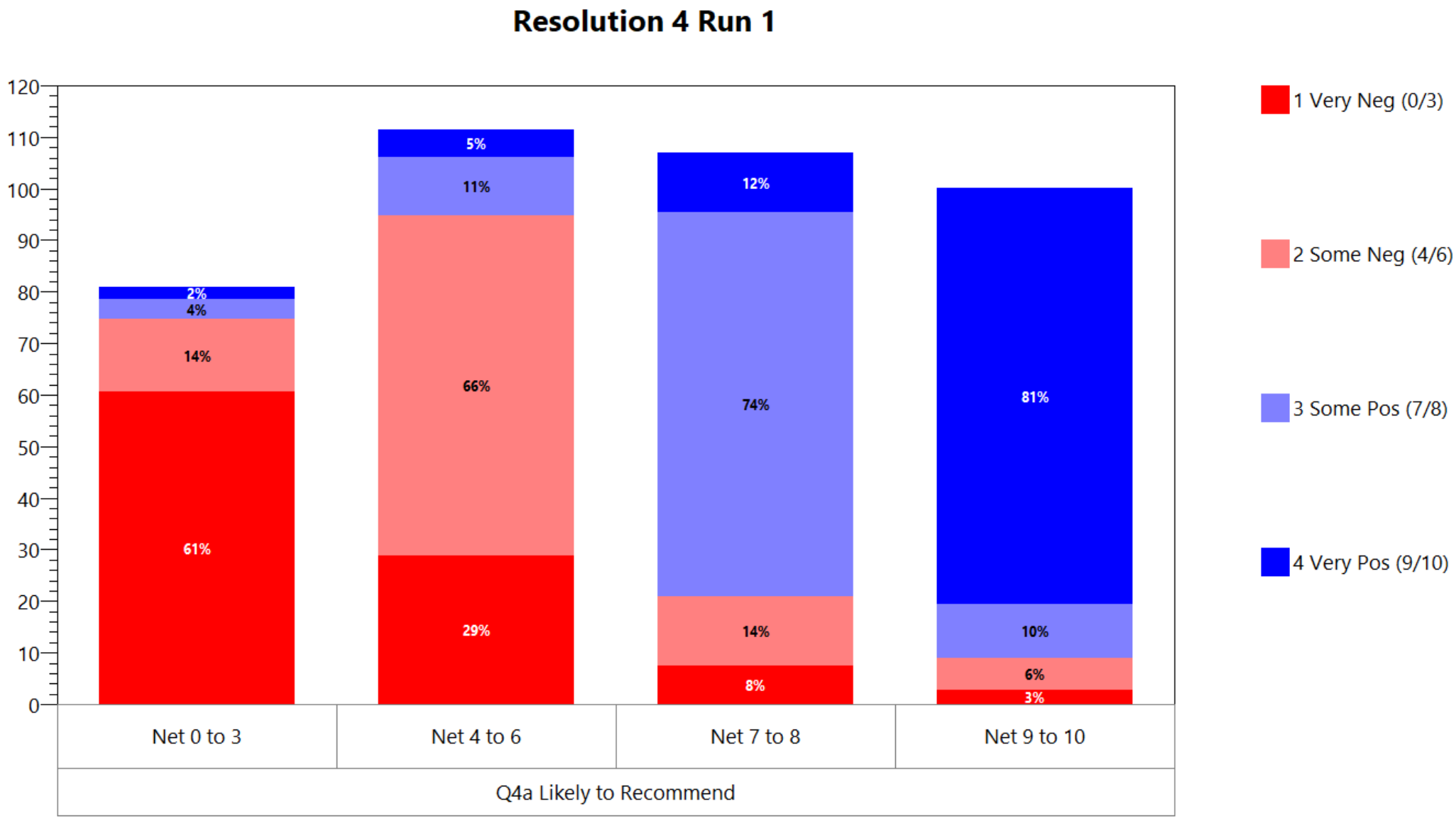
Correctly predicted : 74.63%

Column Percents		Predicted Res4 Run5			
		Very Neg (0/3)	Some Neg (4/6)	Some Pos (7/8)	Very Pos (9/10)
Q4a Likely to Rec	Cases	437	7,972	10,989	9,747
	Net 0 to 3	55%	14%	3%	2%
	Net 4 to 6	32%	65%	10%	6%
	Net 7 to 8	10%	14%	78%	12%
	Net 9 to 10	3%	7%	9%	80%

- Correct 74.26% avg
- Base counts quite disparate (esp. 0/3, max=909, min=93)
- Cells %s no longer identical

But still has a strong diagonal structure

As Chart


Resolution 4 Run 1
120
110
100
90
80
70
60
50
40
30
20
10
0
2%
4%
14%
61%
5%
11%
66%
29%
12%
74%
14%
8%
81%
10%
6%
3%
Net 0 to 3
Net 4 to 6
Net 7 to 8
Net 9 to 10
Q4a Likely to Recommend
1 Very Neg (0/3)
2 Some Neg (4/6)
3 Some Pos (7/8)
4 Very Pos (9/10)

At Highest Resolution (Predicting Q4a itself)

Correctly predicted : 46.86%

Column Percents	Predicted Res11 Run1										
	0	1	2	3	4	5	6	7	8	9	10
Q4a Likely to Recommend: Cases	420	6	17	88	38	4,256	3,633	3,472	6,926	1,833	8,456
0	41%	33%	12%	9%	8%	5%	4%	1%	1%	0%	1%
1	3%	50%	6%	6%		2%	2%	0%	0%	0%	0%
2	2%		53%	5%	5%	3%	3%	1%	1%	0%	1%
3	6%		6%	42%	8%	4%	4%	1%	1%	1%	1%
4	5%			8%	42%	5%	5%	1%	1%	0%	1%
5	18%	17%	12%	22%	11%	44%	26%	4%	4%	2%	3%
6	11%		12%	8%	13%	18%	34%	4%	3%	2%	3%
7	5%				8%	6%	7%	48%	31%	5%	5%
8	6%			1%		7%	9%	31%	50%	10%	8%
9	2%				3%	2%	3%	3%	3%	48%	28%
10	2%				3%	3%	4%	5%	4%	31%	51%

Correctly predicted : 46.50%

Column Percents	Predicted Res11 Run2										
	0	1	2	3	4	5	6	7	8	9	10
Q4a Likely to Recommend: Cases	133	7	43	20	36	4,447	3,000	3,405	8,312	985	8,757
0	54%	29%	9%	10%	14%	7%	4%	2%	1%	0%	1%
1	5%	43%	7%	5%		2%	2%	1%	0%	0%	0%
2	1%		53%	5%	6%	3%	2%	1%	1%	0%	1%
3	4%	14%	5%	60%	8%	4%	4%	1%	1%	1%	1%
4	4%		2%		42%	5%	5%	2%	1%	0%	1%
5	17%	14%	12%	20%	14%	44%	28%	6%	5%	3%	3%
6	6%		7%		11%	19%	35%	6%	4%	3%	3%
7	2%		2%		3%	6%	6%	46%	29%	4%	4%
8	6%					6%	8%	27%	47%	7%	7%
9	1%					2%	3%	3%	4%	54%	29%
10	1%		2%		3%	3%	3%	6%	5%	27%	50%

Correctly predicted : 46.18%

Column Percents	Predicted Res11 Run3										
	0	1	2	3	4	5	6	7	8	9	10
Q4a Likely to Recommend: Cases	399	5	12	16	47	5,727	1,817	6,420	4,767	3,091	6,844
0	41%	20%	8%	19%	11%	6%	2%	1%	1%	1%	1%
1	4%	60%	8%			2%	1%	0%	0%	0%	0%
2	3%		42%	13%	9%	3%	2%	1%	1%	0%	1%
3	6%		17%	50%	6%	5%	3%	1%	1%	1%	1%
4	4%				47%	5%	4%	2%	1%	0%	1%
5	17%	20%		19%	11%	40%	28%	5%	5%	3%	3%
6	12%		17%		13%	21%	37%	5%	4%	2%	3%
7	4%				2%	6%	7%	43%	26%	4%	4%
8	7%					7%	9%	35%	51%	9%	7%
9	2%		8%			2%	3%	3%	4%	45%	26%
10	2%				2%	3%	4%	5%	5%	34%	54%

Correctly predicted : 46.38%

Column Percents	Predicted Res11 Run4										
	0	1	2	3	4	5	6	7	8	9	10
Q4a Likely to Recommend: Cases	124	20	28	15	22	5,041	2,935	2,310	8,938	1,556	8,156
0	55%	15%	11%	13%	18%	7%	4%	2%	1%	0%	1%
1	3%	55%	7%			2%	1%	1%	0%	0%	0%
2	2%		54%	7%	5%	3%	2%	1%	1%	0%	1%
3	5%	15%	11%	53%	5%	5%	4%	1%	1%	1%	1%
4	5%	5%			50%	5%	4%	2%	1%	0%	1%
5	15%	10%	7%	27%		42%	26%	5%	5%	3%	3%
6	6%		7%		14%	19%	35%	5%	4%	2%	2%
7	2%				9%	6%	7%	49%	31%	4%	4%
8	7%					6%	9%	27%	46%	7%	7%
9	1%		4%			2%	3%	2%	4%	51%	28%
10						3%	4%	5%	5%	32%	51%

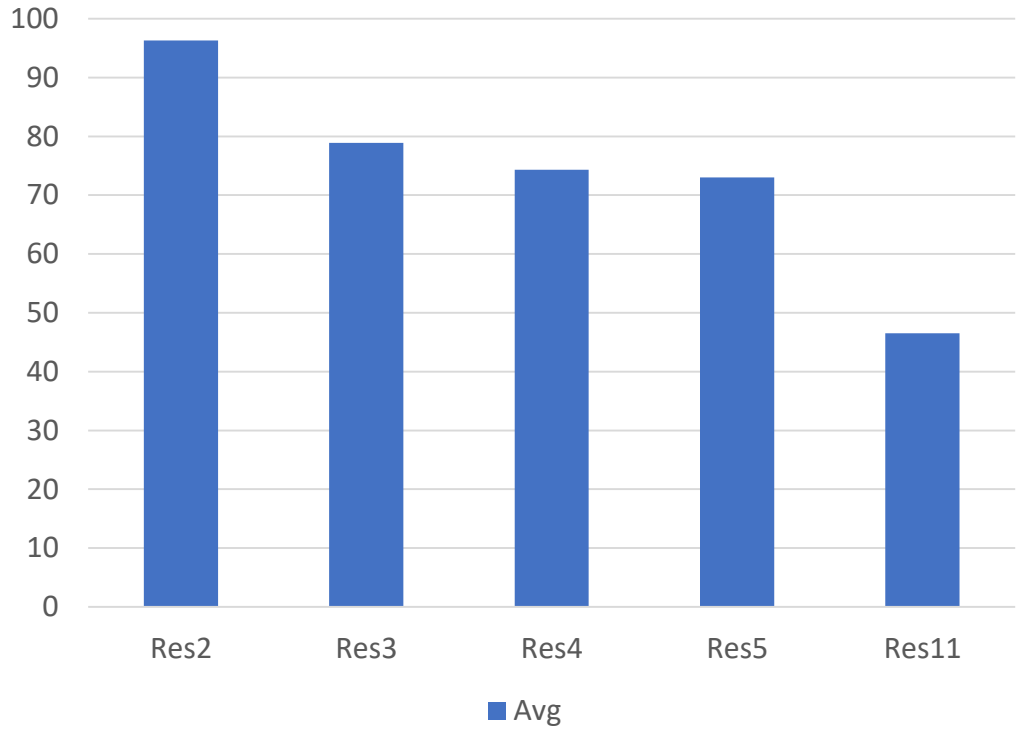
Correctly predicted : 46.42%

Column Percents	Predicted Res11 Run5										
	0	1	2	3	4	5	6	7	8	9	10
Q4a Likely to Recommend: Cases	214	7	13	11	25	3,859	3,607	2,164	9,059	2,455	7,731
0	47%	29%	8%	18%	12%	7%	4%	2%	1%	0%	1%
1	5%	43%	8%			2%	2%	0%	0%	0%	0%
2	3%		62%	9%	8%	3%	3%	1%	1%	0%	1%
3	6%	14%	8%	55%	8%	5%	4%	1%	1%	1%	1%
4	5%				44%	5%	5%	2%	1%	0%	1%
5	17%	14%	8%	18%	12%	46%	27%	5%	5%	3%	3%
6	8%		8%		12%	18%	34%	5%	4%	2%	3%
7	3%					5%	7%	51%	32%	5%	4%
8	5%					5%	8%	25%	46%	9%	7%
9	0%					2%	3%	3%	4%	45%	27%
10	1%				4%	3%	4%	5%	5%	34%	51%

- Now only 46.5% correct
- A lot more noise
- Overlaps on 5/6, 7/8, 9/10
- Base and % disparities

Correct Ratios - All Runs

	Runs					
	1	2	3	4	5	Avg
Res2	96.3	96.31	96.3	96.31	96.31	**96.306**
Res3	78.27	78.84	78.84	79.11	79.31	**78.874**
Res4	74.11	74.24	74	74.42	74.63	**74.28**
Res5	73.05	72.95	72.9	72.93	73.07	**72.98**
Res11	46.86	46.5	46.18	46.38	46.42	**46.468**



- Ratios at each resolution are stable across all runs
- But at higher resolutions, frequencies and percents increasingly diverge

Is resolving to 11 points (avg 46.5) too ambitious?

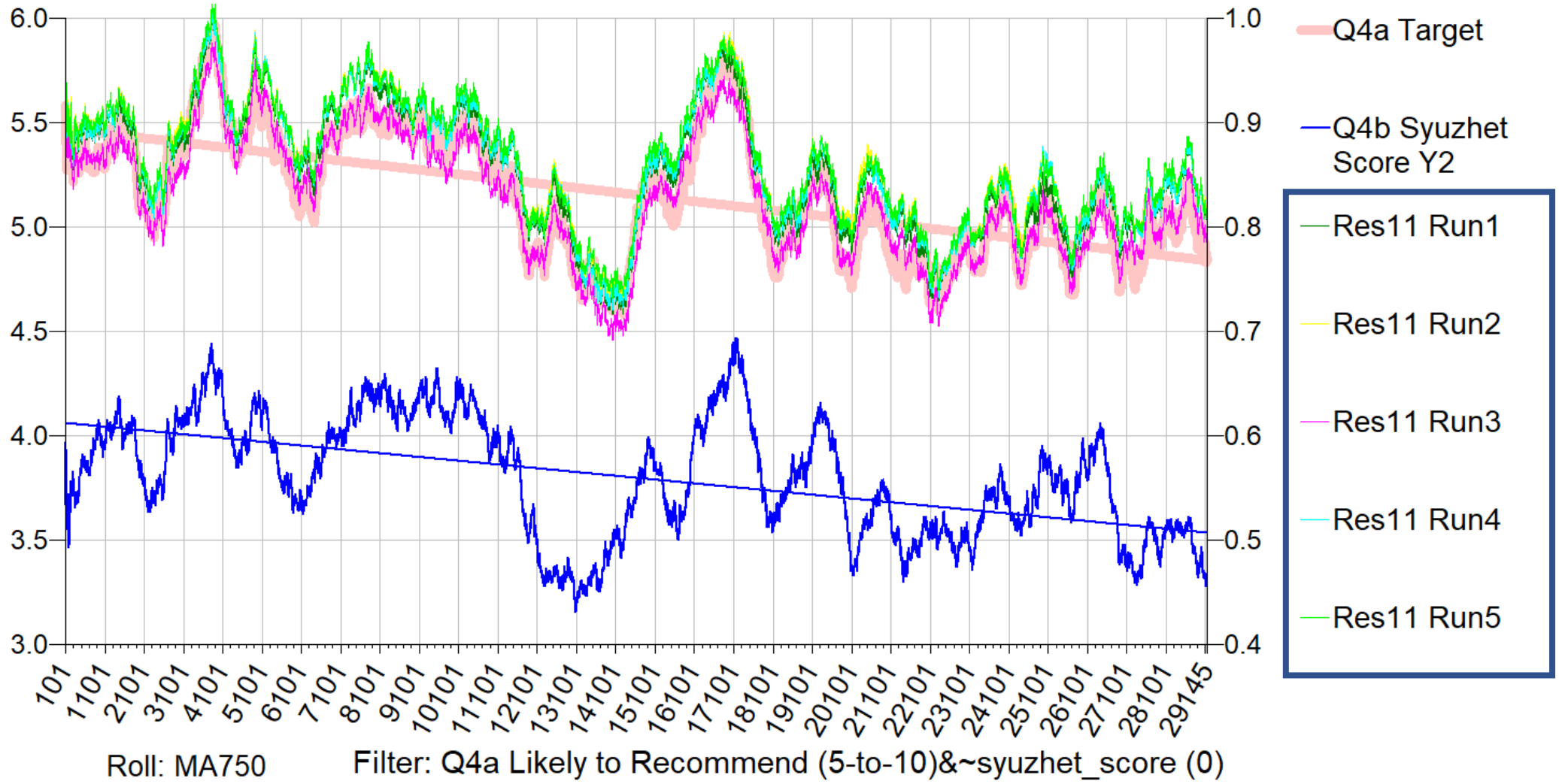
Remove Noise

- Ignore verbatims with syuzhet score = 0
  - Not enough meat for syuzhet to register, so ML will also have difficulties
- Q4a frequency counts are quite low for ratings 0 to 4
- Respondent engagement (as a considered verbatim response) loosely corresponds with rating
  - longer, detailed, expansive for higher ratings, but short or irrelevant for lower
  - Small sample sizes and low engagement responses will distract ML

So filter to syuzhet_score<>0 and Q4a is a 5 to 10

As Chart

- Predictions run a little high, but
- Better match than syuzhet

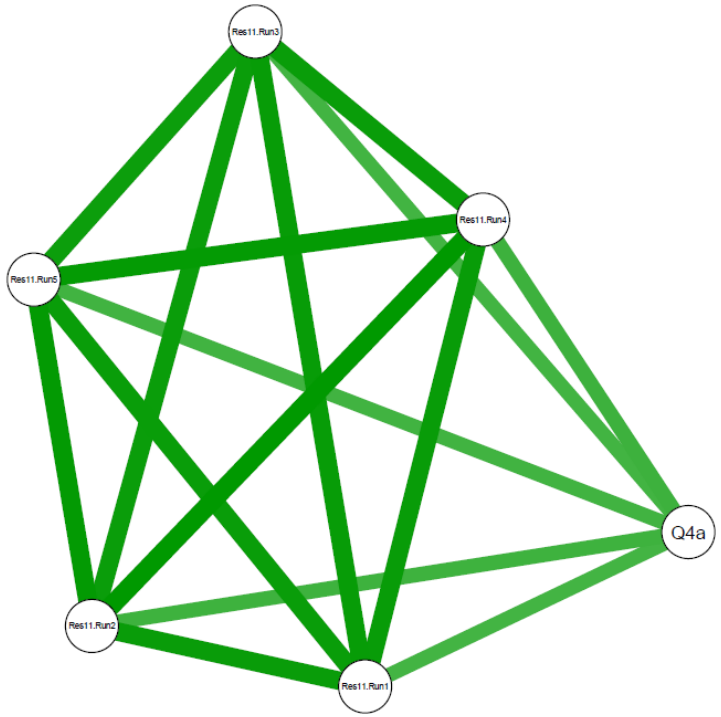


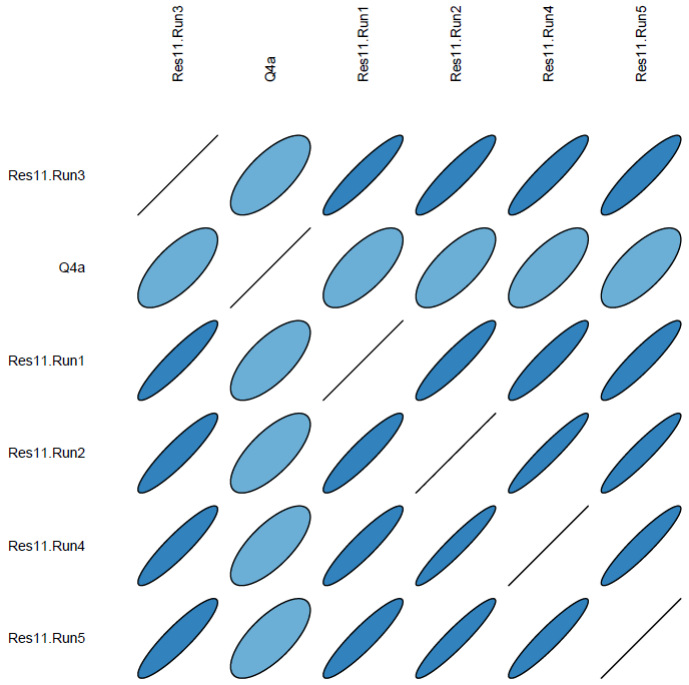
As Per Case Correlation: 0.71

Filter: Q4a Likely to Recommend (5-to-10)&~syuzhet_score (0)

Frequencies	Q4a	Res11 Run1	Res11 Run2	Res11 Run3	Res11 Run4	Res11 Run5
CaseSeq 1	8	8	8	7	8	8
2	7	8	8	8	8	8
3	7	8	8	8	8	8
4	8	7	7	7	7	7
6	7	6	7	7	6	6
7	6	5	5	5	5	5
9	10	10	10	10	10	10
11	10	10	10	10	10	10
14	9	10	10	10	10	10
15	10	10	10	10	10	10
16	9	9	10	9	10	10
17	6	6	8	7	8	8
18	10	10	10	10	10	10
19	10	9	10	9	9	9
20	10	10	10	10	10	10
21	7	8	8	8	8	8
23	10	10	10	10	10	10
24	7	7	7	7	7	8
25	7	7	8	7	8	8
26	9	5	5	5	5	5
28	8	6	5	5	6	6
30	7	10	10	10	10	10
31	9	8	8	8	8	8
32	10	9	9	9	9	9
33	6	6	6	6	6	6
34	9	7	7	7	7	7
35	8	8	8	8	8	8
36	7	7	7	7	8	8
38	10	10	10	10	10	10

	Q4a	Res11.Run1	Res11.Run2	Res11.Run3	Res11.Run4	Res11.Run5
Q4a	1	0.701	0.72	0.704	0.725	0.711
Res11.Run1	0.701	1	0.917	0.925	0.919	0.917
Res11.Run2	0.72	0.917	1	0.912	0.949	0.938
Res11.Run3	0.704	0.925	0.912	1	0.916	0.903
Res11.Run4	0.725	0.919	0.949	0.916	1	0.933
Res11.Run5	0.711	0.917	0.938	0.903	0.933	1





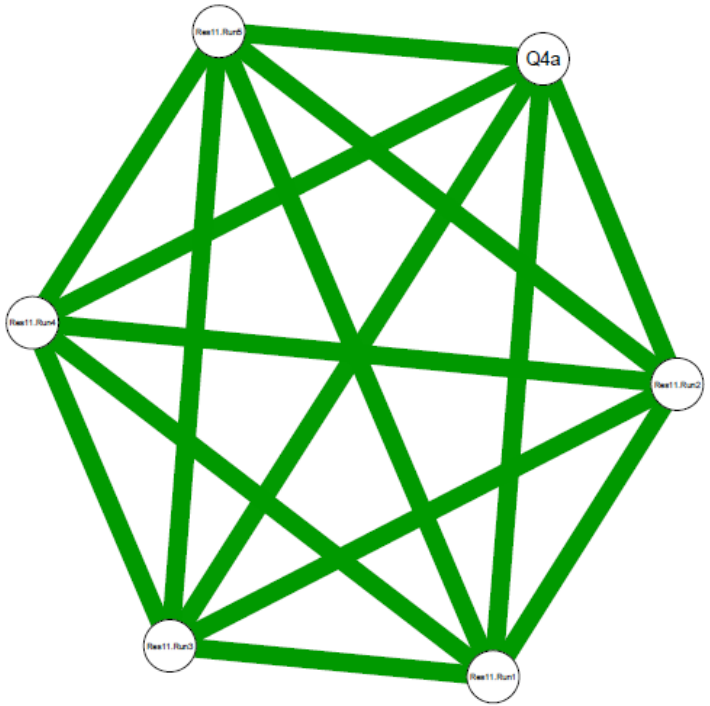
As Per Verbatim Correlation: 0.9924

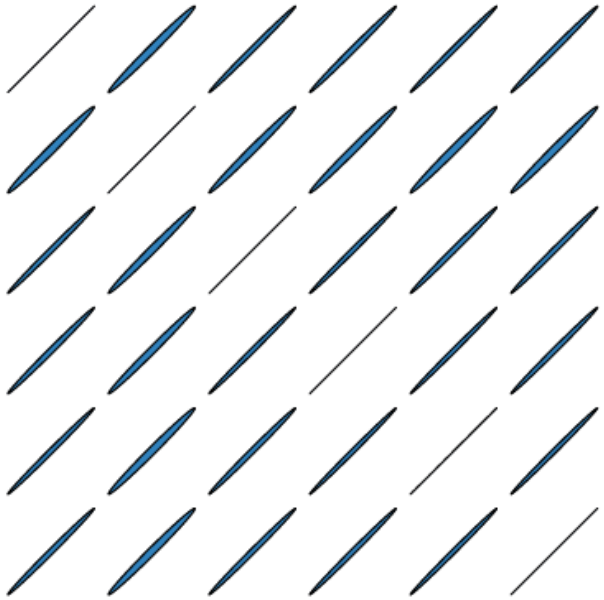
Filter: Q4a Likely to Recommend (5-to-10)&~syuzhet_score (0)

Frequencies

Q4b	Q4a	Res11 Run1	Res11 Run2	Res11 Run3	Res11 Run4	Res11 Run5
100% refund	8	8	8	7	8	8
AFTER 60 YEARS WITH THEM AND NO	10	10	10	10	10	10
All good	8	9	10	9	10	9
Always give good service, no hassles	10	10	10	10	10	10
always used them never had any probl	10	9	9	9	9	9
best service best fees least gaps best in	9	10	10	10	10	10
better packages for all insurance and di	7	8	8	7	8	8
better rates - less hassle	7	7	7	7	7	7
Better Rebates	129	126	126	126	144	126
Better use of limits on benefits. Being	7	7	7	7	7	7
cheaper ins and better payouts	8	8	8	7	8	8
Cheaper, more personalized insurance	7	7	7	7	7	7
claims are fast, good range of benefits	9	10	10	10	10	10
Cove is what I need and have never ha	9	10	10	9	10	10
Cover and their gap saver	9	10	10	10	10	10
Covers dental, physio & medical that	10	7	8	7	8	8
decrease the prices of the premiums	7	8	8	7	8	8
Does not have preferred services which	6	8	8	7	8	8
don't have any reason	5	5	5	5	5	5
don't recommend products to anyone	5	5	5	5	5	5
Don't recommend any thing	6	5	5	5	5	5
Don't talk about it	11	10	10	10	10	10
Don't use everything that is offered. Fill	6	6	6	5	6	6
Drop their premiums. Pay higher refun	8	8	8	8	8	8
Easy and friendly to deal with.	9	10	10	10	10	10
Easy to claim, good levels of options	9	9	9	9	9	9
Easy to understand, easy to claim.	9	10	10	9	10	9
excellent service	133	140	140	140	140	140
Excellent service and that they are a Go	10	10	10	10	10	10

	Q4a	Res11.Run1	Res11.Run2	Res11.Run3	Res11.Run4	Res11.Run5
Q4a	1	0.993	0.993	0.992	0.992	0.992
Res11.Run1	0.993	1	0.997	0.997	0.996	0.996
Res11.Run2	0.993	0.997	1	0.996	0.998	0.997
Res11.Run3	0.992	0.997	0.996	1	0.997	0.996
Res11.Run4	0.992	0.996	0.998	0.997	1	0.997
Res11.Run5	0.992	0.996	0.997	0.996	0.997	1





Conclusions:

- ML is a better predictor than the syuzhet weighted words approach
- Consistent *Correct* ratios across runs
- Resolving to 11 points is the practical limit (on this test set)
- Excellent match on de-noised per-case plots
- De-noised correlation > 0.99 (by unique verbatim strings)

Therefore: The model can be deemed to be sufficiently valid to deploy against unseen cases