Data 410 Project Rough Draft

Daniel Krasnov, Keiran Malott, Ross Cooper

2023-04-07

Contents

Guidelines

- An introduction to the dataset, and the scientific hypotheses you will investigate.
- A descriptive analysis of the data; give a detailed description about the data including the number of variables, variables types, summary statistics, graphs of data, etc..

– A regression analysis that addresses your scientific hypothesis, using all the regression model building techniques you have learned. Model and data appropriateness diagnostics are expected. Plots and tables are highly encouraged, where you need to include the interpretation for each plot/table.

– Conclusions and recommendations: give your conclusion based on your regression analysis such as important variables identified, the most proper regression model you have discovered, how your regression assumptions may be violated and how they affect your results, etc

Data Collection Method

describe how we scappred Not How Girls Work Subreddit

Variable Description

Show example of our full data set show

Data Preprocessing

Explain we only used Sarcastic and body columns. We used regex to remove /s. We then remove stopwords, lemmatize, etc.

Feature Extraction Methods

interpreting text as data is lots of work say something about why that's tough and we need ways to vectorize text.

TF-IDF

Explain tf-idf. Show what final result is for design matrix and explain how you got there.

Word2Vec

Explain word2vec. Show what final result is for design matrix and explain how you got there.

Glove

Explain Glove. Show what final result is for design matrix and explain how you got there.

Evaluation Metrics

Describe Precision, Recall, F1.

Regression Analysis

This analysis is a comparison of logistic model performance when using 3 types of feature extraction.

Daniel's Analysis (better title later)

Base Model

The dimension of the DTM is too large by default. Can't use bestGLM as too many predictors. Do form of best subset by checking quantities of TF-IDF.

The 95% percentile gave the best AIC so this is the base model I will select.

There are several NAs, not good. Let's get results.

High VIFs, bad need to reduce collineairty. We try PCA now.

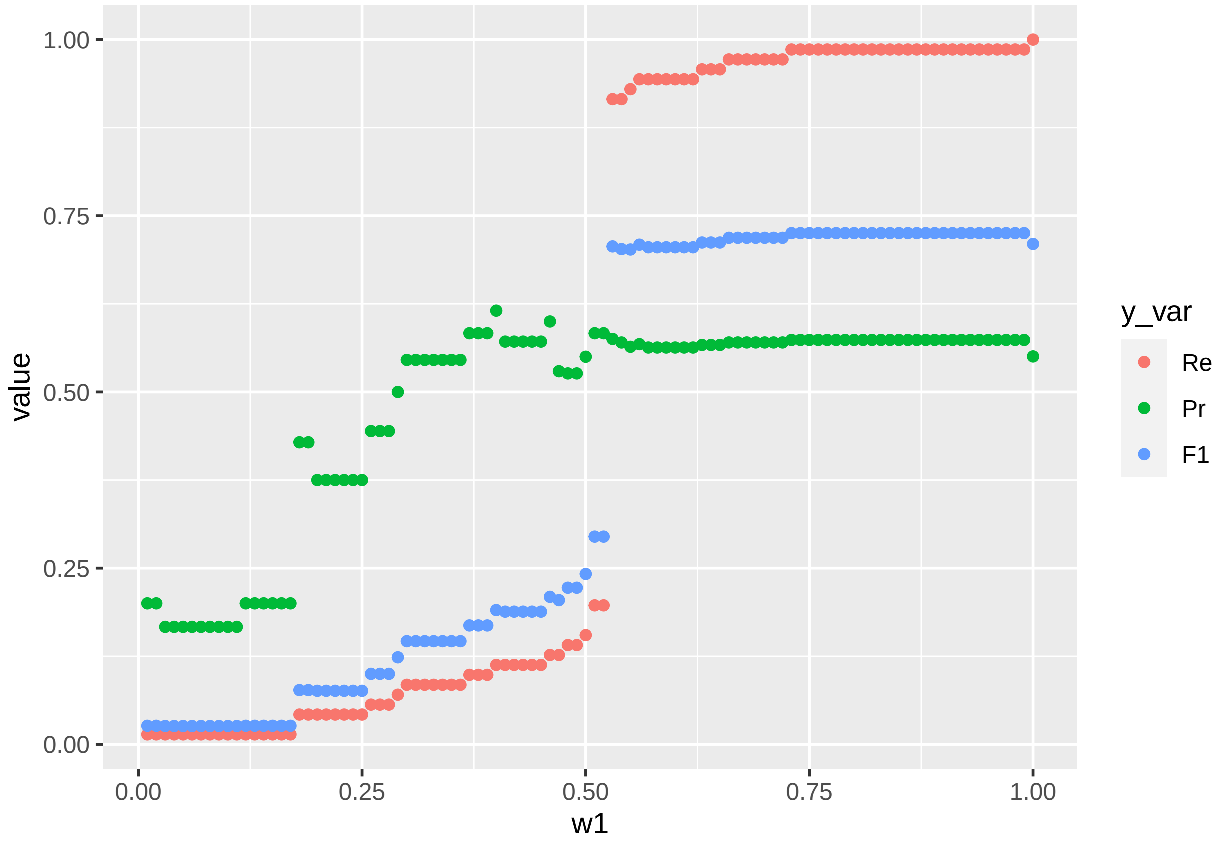
PCA

We keep up to 90th percentile

Very not good predicts everything as not sar.

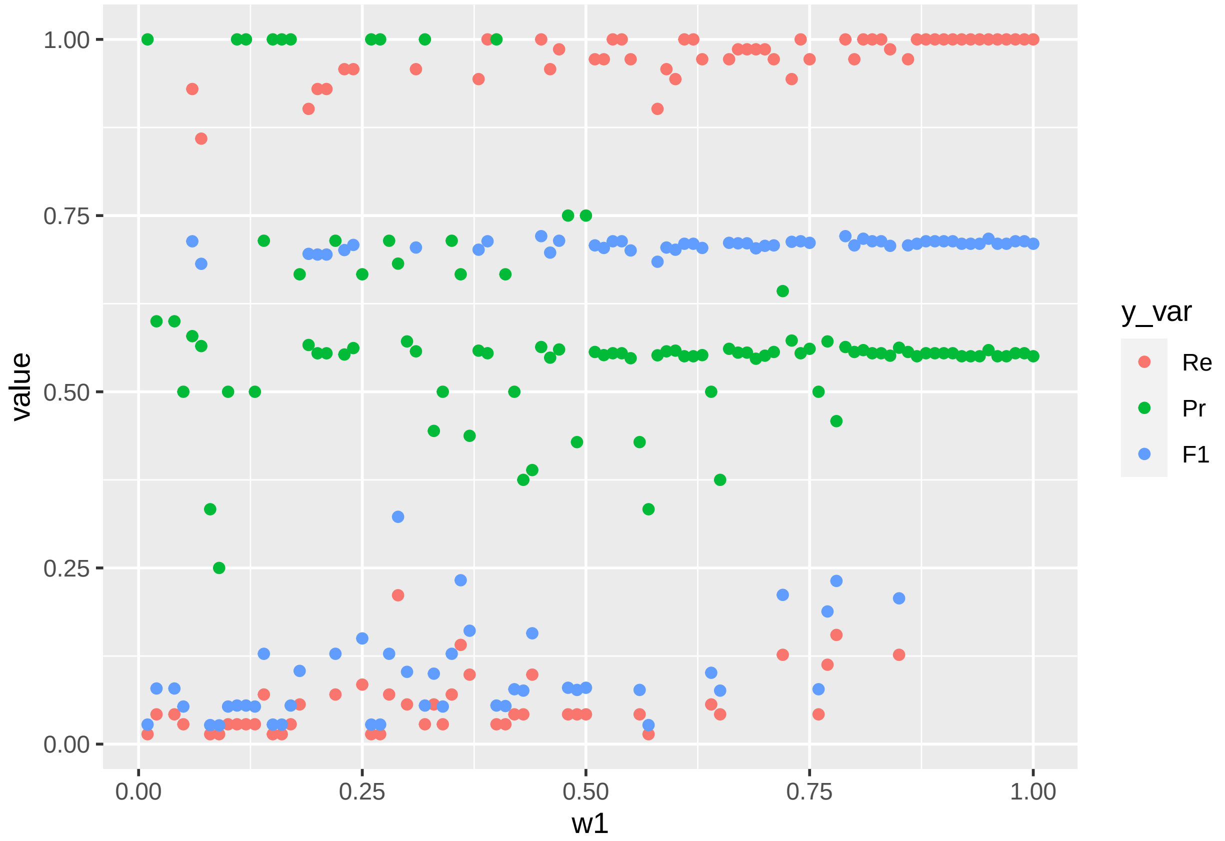
LASSO

Weighted Base Model



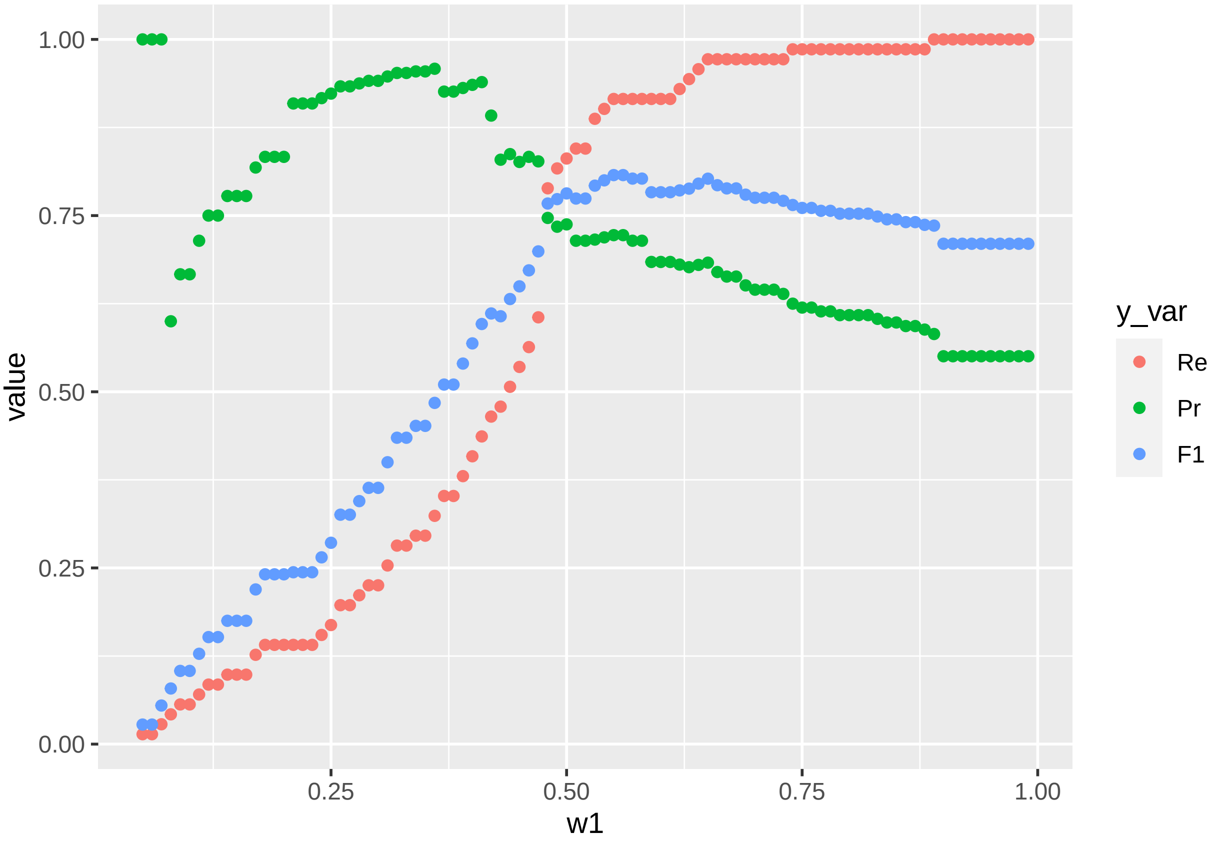
We want a F1 that balances Recall and Precision but still focuses on Recall. Anything around 0.75 is good.

Weighted PCA



F1 holds pretty steady. We go for the one that has the highest Recall, a reasonable precision, and w good F1. 0.62 has

Weighted LASSO



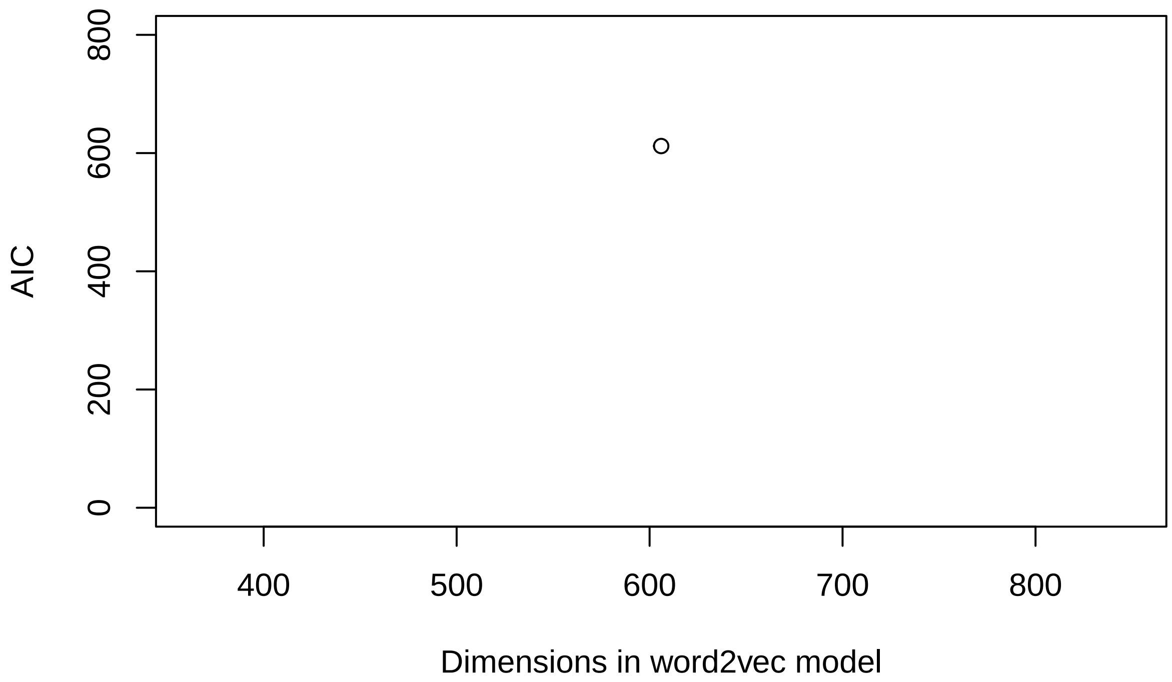
0.55 is best

Keiran's Analysis (better title later)

Ross' Analysis (better title later)

```
## [1] FALSE
```

```
## INFO  [14:44:28.181] epoch 1, loss 0.2111
## INFO  [14:44:28.250] epoch 2, loss 0.0802
## INFO  [14:44:28.274] epoch 3, loss 0.0410
## INFO  [14:44:28.285] epoch 4, loss 0.0224
## INFO  [14:44:28.295] epoch 5, loss 0.0149
## INFO  [14:44:28.305] epoch 6, loss 0.0105
## INFO  [14:44:28.315] epoch 7, loss 0.0077
## INFO  [14:44:28.325] epoch 8, loss 0.0058
## INFO  [14:44:28.335] epoch 9, loss 0.0044
## INFO  [14:44:28.345] epoch 10, loss 0.0034
```



140 dimensions gives the best model for logistic regression

Base Model

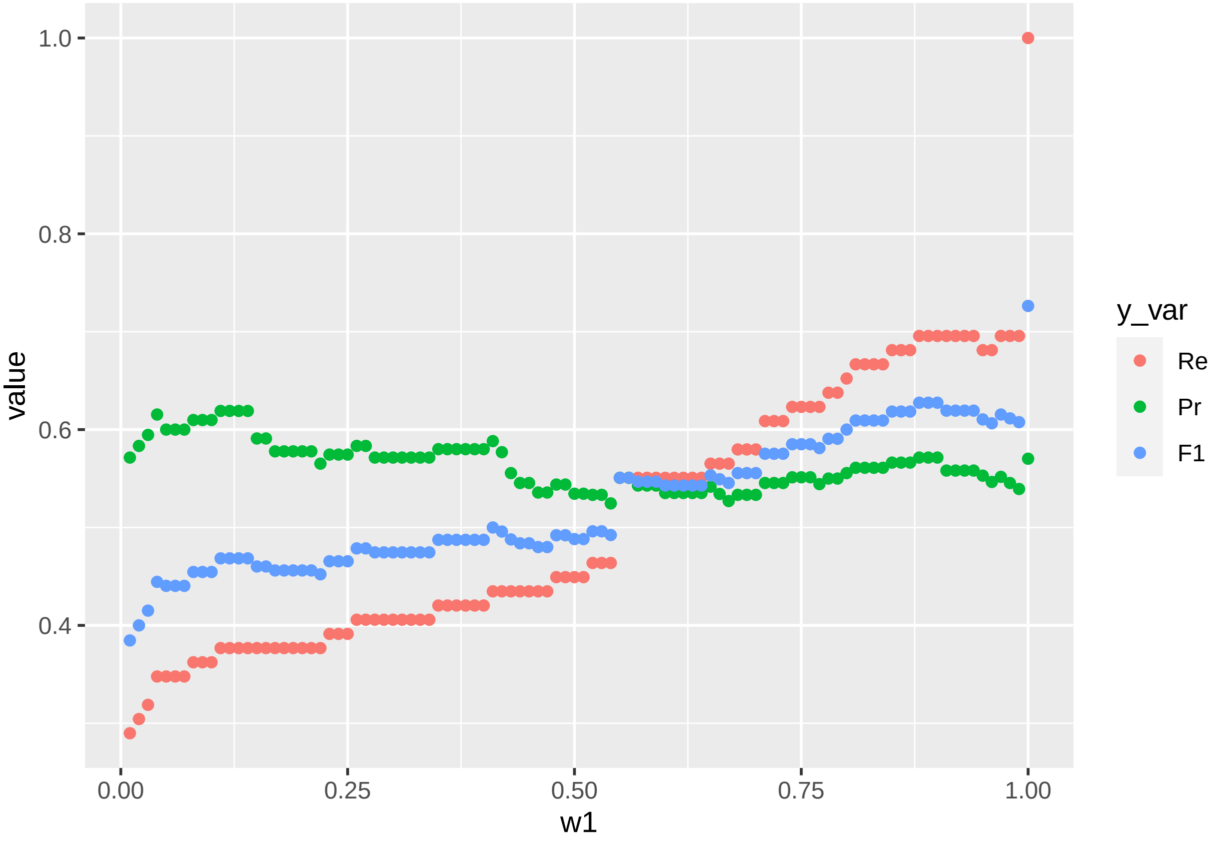
```
## INFO  [14:44:35.295] epoch 1, loss 0.2148
## INFO  [14:44:35.308] epoch 2, loss 0.0827
## INFO  [14:44:35.319] epoch 3, loss 0.0428
## INFO  [14:44:35.330] epoch 4, loss 0.0223
## INFO  [14:44:35.340] epoch 5, loss 0.0151
## INFO  [14:44:35.350] epoch 6, loss 0.0107
## INFO  [14:44:35.360] epoch 7, loss 0.0078
## INFO  [14:44:35.370] epoch 8, loss 0.0059
## INFO  [14:44:35.381] epoch 9, loss 0.0045
## INFO  [14:44:35.391] epoch 10, loss 0.0034
```

```
##
## Call:
## glm(formula = sarcasm ~ ., family = "binomial", data = training_mat)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -3.04352  -0.56494   0.00741   0.56146   2.32416
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.18132    0.24306  -0.746 0.455672
## X2          -0.80430    0.39803  -2.021 0.043313 *
## X3           0.26995    0.35651   0.757 0.448925
## X4          -0.19573    0.35146  -0.557 0.577586
## X5           0.08480    0.36562   0.232 0.816592
## X6           0.10295    0.29554   0.348 0.727580
## X7          -0.19246    0.30070  -0.640 0.522147
## X8           0.30256    0.33745   0.897 0.369925
## X9          -0.35079    0.41846  -0.838 0.401872
## X10          0.35862    0.42191   0.850 0.395319
## X11         -0.23722    0.40880  -0.580 0.561726
## X12         -0.02709    0.37880  -0.072 0.942983
## X13         -0.06157    0.39653  -0.155 0.876609
## X14         -0.32006    0.34807  -0.920 0.357813
## X15         -0.22650    0.30850  -0.734 0.462825
## X16          0.12119    0.38767   0.313 0.754579
## X17         -0.32444    0.36200  -0.896 0.370115
## X18         -0.39291    0.34599  -1.136 0.256113
## X19         -0.18408    0.42205  -0.436 0.662728
## X20         -0.35601    0.38881  -0.916 0.359845
## X21          0.17480    0.41451   0.422 0.673247
## X22         -0.44739    0.35263  -1.269 0.204534
## X23          0.38562    0.38506   1.001 0.316607
## X24         -0.05360    0.33050  -0.162 0.871162
## X25         -0.39965    0.34612  -1.155 0.248228
## X26         -0.02965    0.36748  -0.081 0.935688
## X27         -0.35053    0.37416  -0.937 0.348836
## X28          0.19979    0.40218   0.497 0.619354
## X29          0.03704    0.39658   0.093 0.925596
## X30          0.33211    0.36395   0.913 0.361486
## X31         -0.33020    0.33152  -0.996 0.319245
## X32          1.82120    0.47636   3.823 0.000132 ***
## X33         -0.55228    0.38969  -1.417 0.156416
## X34         -0.54054    0.37680  -1.435 0.151415
## X35         -0.02912    0.33197  -0.088 0.930094
## X36         -1.19091    0.41674  -2.858 0.004268 **
## X37         -0.03700    0.34159  -0.108 0.913754
## X38          0.09343    0.30960   0.302 0.762813
## X39          0.69047    0.40522   1.704 0.088390 .
## X40         -1.58230    0.40422  -3.914 9.06e-05 ***
## X41          0.79758    0.39966   1.996 0.045972 *
## X42          1.28535    0.42618   3.016 0.002561 **
## X43          0.48205    0.36026   1.338 0.180873
## X44          0.79832    0.35152   2.271 0.023146 *
## X45          0.28660    0.35848   0.800 0.423999
## X46         -0.17861    0.30942  -0.577 0.563766
## X47         -0.38252    0.32640  -1.172 0.241226
## X48         -0.36603    0.37163  -0.985 0.324653
## X49         -1.27813    0.37435  -3.414 0.000640 ***
## X50          0.34549    0.34246   1.009 0.313041
## X51          1.31799    0.42229   3.121 0.001802 **
## X52         -0.22564    0.40439  -0.558 0.576858
## X53         -0.37141    0.40993  -0.906 0.364911
## X54         -0.38636    0.35593  -1.086 0.277700
## X55          1.20889    0.39704   3.045 0.002329 **
## X56         -0.25129    0.36340  -0.692 0.489242
## X57          0.09541    0.36945   0.258 0.796210    
## X58          0.24091    0.36280   0.664 0.506673    
## X59          1.43855    0.41960   3.428 0.000607 ***
## X60          0.16404    0.41465   0.396 0.692402    
## X61         -0.09198    0.32486  -0.283 0.777058    
## X62          0.16990    0.34402   0.494 0.621404    
## X63         -0.17688    0.32088  -0.551 0.581477    
## X64         -0.44214    0.32713  -1.352 0.176519    
## X65          0.38622    0.33337   1.159 0.246639    
## X66          0.67504    0.35749   1.888 0.058989 .  
## X67          0.21146    0.34759   0.608 0.542945    
## X68         -0.08475    0.37750  -0.225 0.822360    
## X69         -0.39300    0.41993  -0.936 0.349343    
## X70          0.51355    0.30222   1.699 0.089271 .  
## X71         -0.30822    0.33081  -0.932 0.351480    
## X72          0.25259    0.42541   0.594 0.552674    
## X73          0.25831    0.37863   0.682 0.495091    
## X74          0.26482    0.36846   0.719 0.472312    
## X75          0.36236    0.35266   1.027 0.304188    
## X76          0.49273    0.35104   1.404 0.160433    
## X77         -0.36528    0.36521  -1.000 0.317218    
## X78          0.31988    0.36744   0.871 0.383998    
## X79         -0.31661    0.35557  -0.890 0.373238    
## X80         -0.99030    0.40393  -2.452 0.014220 *  
## X81          0.90216    0.37599   2.399 0.016422 *  
## X82          0.09909    0.36127   0.274 0.783856    
## X83         -0.78644    0.39043  -2.014 0.043978 *  
## X84         -0.72145    0.41699  -1.730 0.083603 .  
## X85          0.82376    0.36651   2.248 0.024603 *  
## X86         -0.16027    0.39144  -0.409 0.682226    
## X87          0.18701    0.33399   0.560 0.575535    
## X88         -0.01790    0.33022  -0.054 0.956769    
## X89          0.46096    0.32696   1.410 0.158592    
## X90         -0.59128    0.37705  -1.568 0.116839    
## X91         -0.39492    0.33352  -1.184 0.236374    
## X92          0.08287    0.37299   0.222 0.824165    
## X93         -0.73566    0.40562  -1.814 0.069729 .  
## X94         -0.42477    0.31800  -1.336 0.181625    
## X95         -0.81399    0.40117  -2.029 0.042455 *  
## X96          0.03961    0.35693   0.111 0.911627    
## X97         -0.08879    0.37186  -0.239 0.811283    
## X98          0.69703    0.38287   1.821 0.068679 .  
## X99         -0.34844    0.32146  -1.084 0.278402    
## X100        -1.15209    0.42811  -2.691 0.007121 ** 
## X101        -0.13666    0.36018  -0.379 0.704363    
## X102         0.48785    0.36663   1.331 0.183318    
## X103         0.20398    0.40640   0.502 0.615730    
## X104        -0.90009    0.37768  -2.383 0.017162 *  
## X105         0.08395    0.41682   0.201 0.840376    
## X106        -0.28358    0.37671  -0.753 0.451584    
## X107        -0.22330    0.41212  -0.542 0.587931    
## X108        -0.38536    0.36608  -1.053 0.292496    
## X109         0.35271    0.39543   0.892 0.372419    
## X110        -0.10138    0.34440  -0.294 0.768485    
## X111         0.17131    0.35199   0.487 0.626473
## X112         0.67193    0.38042   1.766 0.077350 .
## X113        -0.30665    0.37494  -0.818 0.413443
## X114         0.90209    0.35505   2.541 0.011062 *
## X115         0.48582    0.34830   1.395 0.163063
## X116         0.11279    0.38057   0.296 0.766951
## X117         0.40941    0.36850   1.111 0.266556
## X118        -0.06746    0.36721  -0.184 0.854240
## X119         1.11209    0.42199   2.635 0.008405 **
## X120        -0.33632    0.39231  -0.857 0.391290
## X121         0.28905    0.37868   0.763 0.445282
## X122        -0.84912    0.37576  -2.260 0.023837 *
## X123        -0.54788    0.39212  -1.397 0.162344
## X124        -0.15227    0.39417  -0.386 0.699268
## X125         0.39521    0.40522   0.975 0.329416
## X126         0.41227    0.37799   1.091 0.275405
## X127         0.31411    0.40164   0.782 0.434177
## X128         0.24962    0.37025   0.674 0.500192
## X129        -0.23264    0.36264  -0.642 0.521187
## X130         0.33471    0.33248   1.007 0.314071
## X131        -0.40419    0.35094  -1.152 0.249418
## X132         0.82590    0.38356   2.153 0.031298 *
## X133        -0.01843    0.39813  -0.046 0.963084
## X134        -0.12506    0.41360  -0.302 0.762371
## X135        -0.55975    0.39332  -1.423 0.154692
## X136         0.11077    0.32941   0.336 0.736675
## X137        -0.36117    0.38765  -0.932 0.351501
## X138         0.08445    0.34777   0.243 0.808138
## X139         0.17834    0.34322   0.520 0.603333
## X140        -0.42222    0.40086  -1.053 0.292212
## X141        -0.20392    0.39037  -0.522 0.601405
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 672.19  on 484  degrees of freedom
## Residual deviance: 358.44  on 344  degrees of freedom
## AIC: 640.44
##
## Number of Fisher Scoring iterations: 7
```

PCA

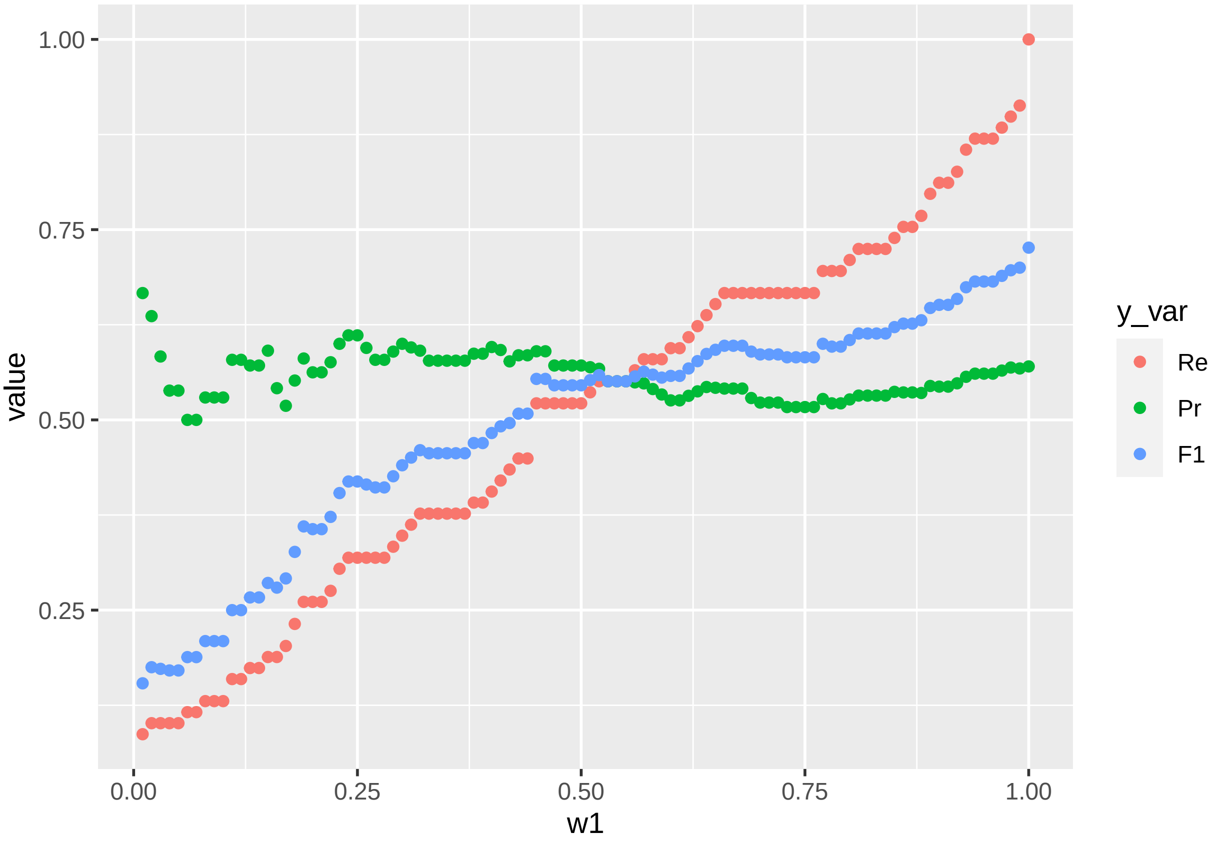
LASSO

Weighted Base Model



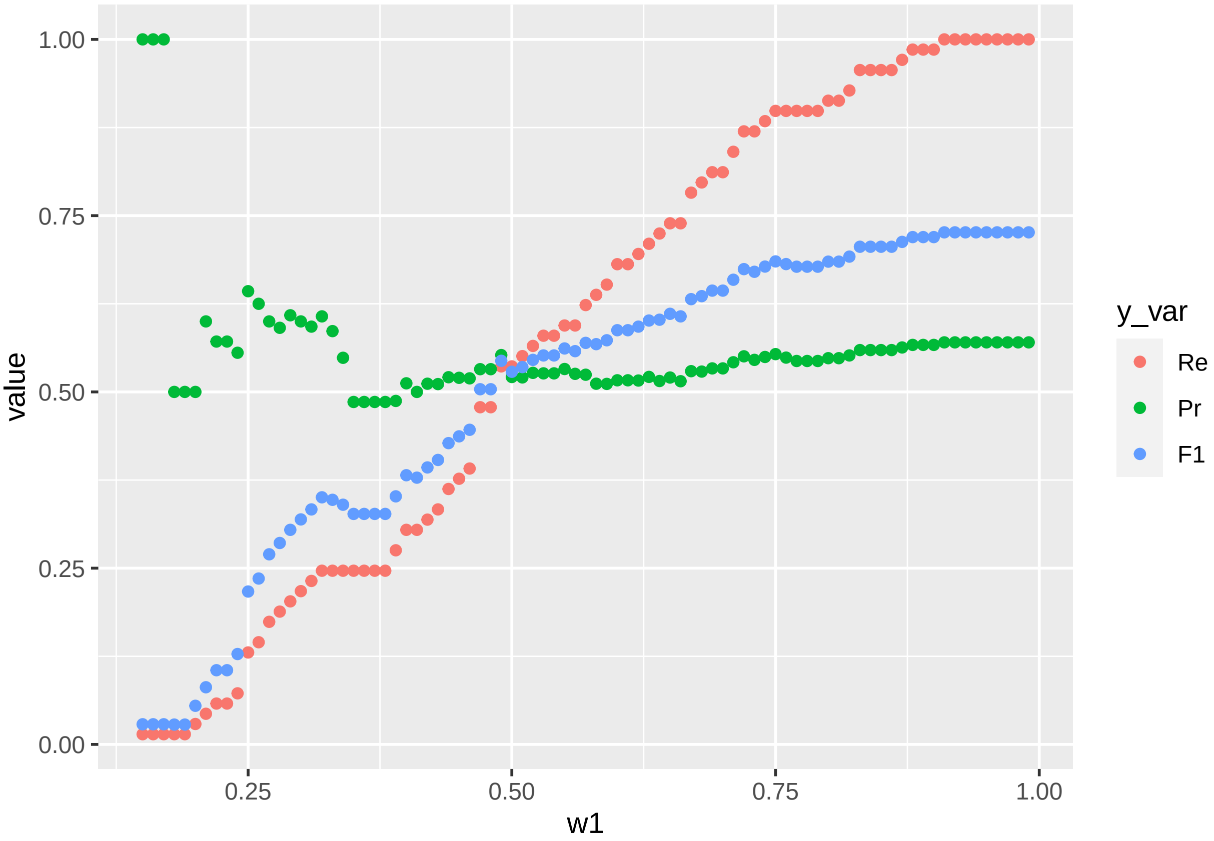
We want a F1 that balances Recall and Precision but still focuses on Recall. Anything around 0.7 is good.

Weighted PCA



Precision holds pretty steady while recall and F1 increase with weighting. We want a weighting with a good balance of all three where the recall plateaus. All scores are good at 0.9, which is also the beginning of the plateau.

Weighted LASSO



Recall and F1 plateau at 0.9