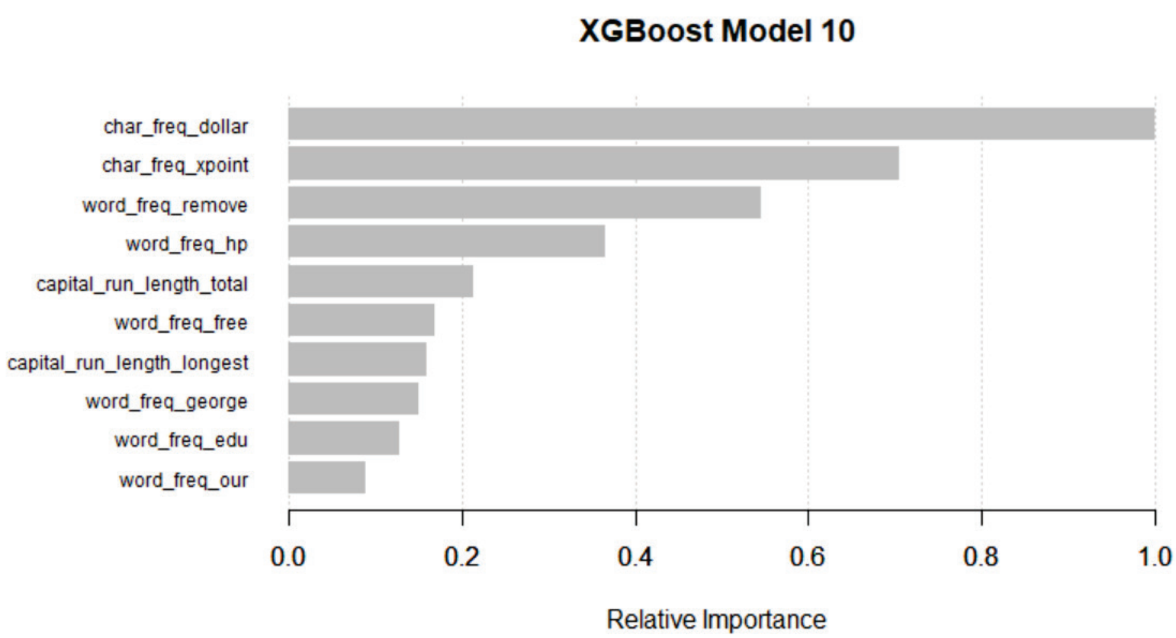
Assignment 8 Joel Riesen

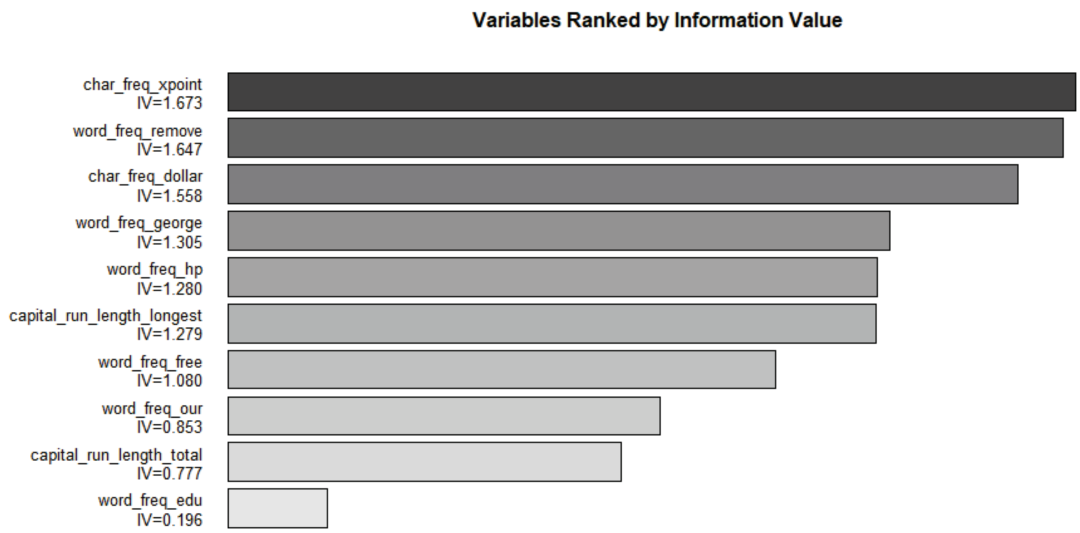
Computational Exploratory Data Analysis

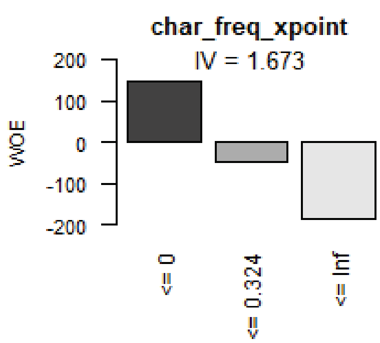
We will start the process to find out which predictor variables are essential. To help us pair down the data and make the process more efficient. Looking at the spam.df, it has 61 variables and 4601 observations.

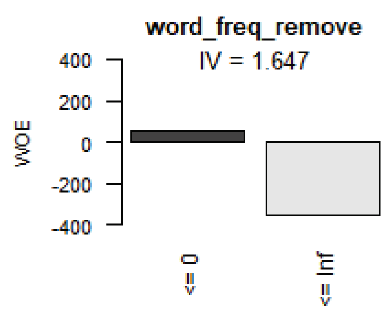
We will use the the XGBoost tree model to start us off. We will be ising the settings of max_depth = 4 and nrounds = 10. Using this information we get the the Top 10 :[1] "char_freq_dollar", [2] "char_freq_xpoint", [3] "word_freq_remove" , [4] "word_freq_hp", [5] "capital_run_length_total" , [6] "word_freq_free" , [7] "capital_run_length_longest", [8] "word_freq_george" ,[9] "word_freq_edu", [10] "word_freq_our"
The training errors for them
[1] train-error:0.086281
[2] train-error:0.079379
[3] train-error:0.078947
[4] train-error:0.070751
[5] train-error:0.064711
[6] train-error:0.058240
[7] train-error:0.056083
[8] train-error:0.055220
[9] train-error:0.053063
[10] train-error:0.050906
I think it is pretty good that doing just 10 boosting iterations that it moves from .08 to a .05. Using the same max_depth = 4 but changing the nrounds to 20 the last iteration is now at 0.03537. Not as big of a move as before but still pretty good.

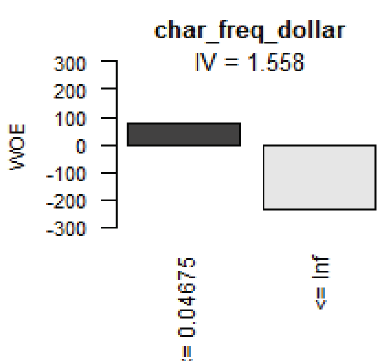


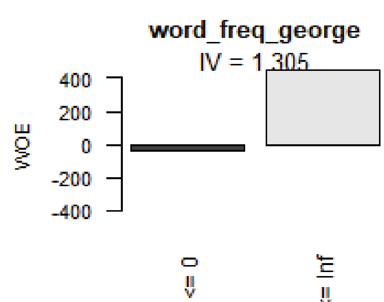
We are continuing with the top ten predictor variables from the XGBoost model and then using the WOE transformation.

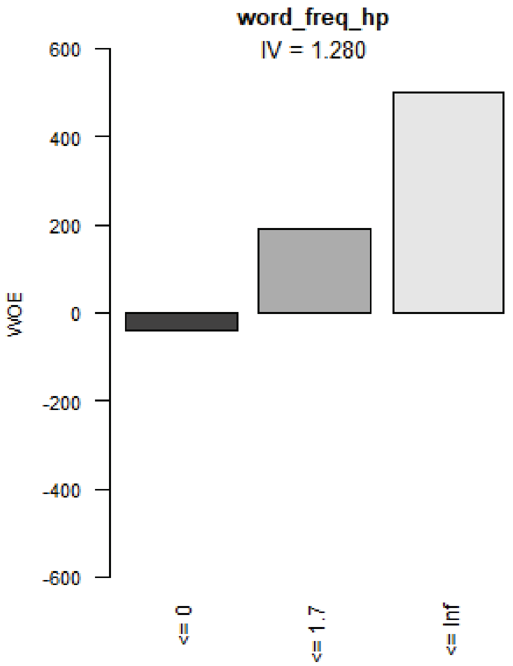


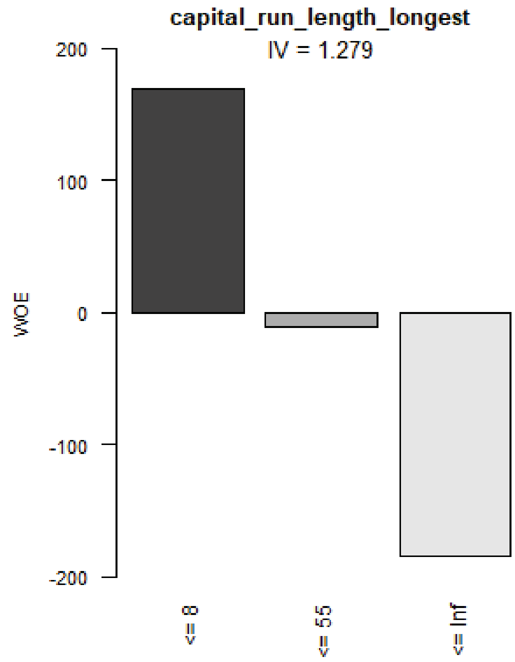


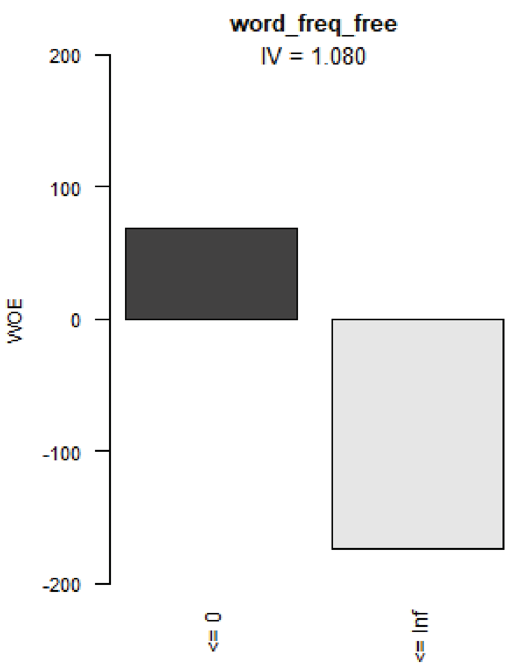


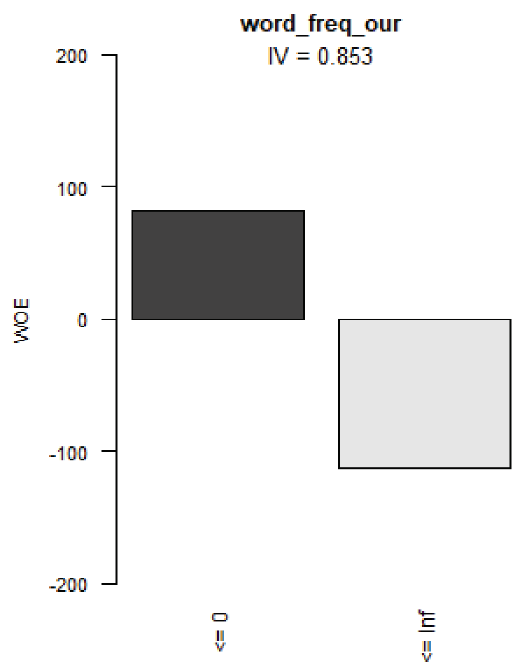


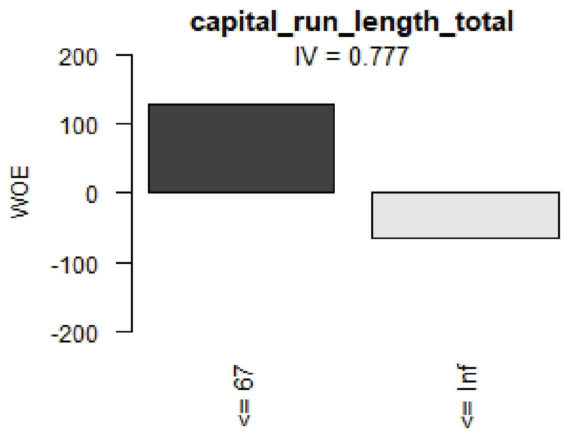


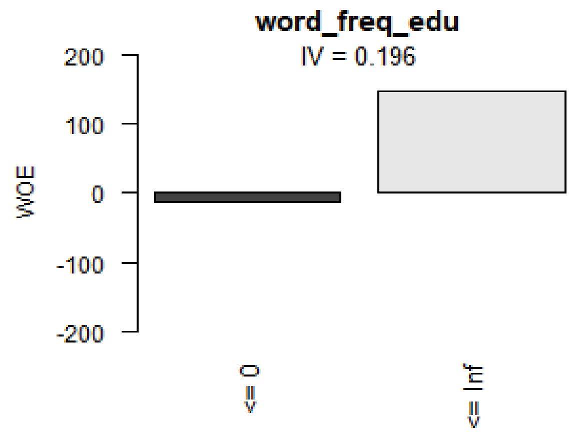

word_freq_hp
IV = 1.280
WOE
600
400
200
0
-200
-400
-600
<= 0
<= 1.7
<= Inf


capital_run_length_longest
IV = 1.279
WOE
200
100
0
-100
-200
<= 8
<= 55
<= Inf


word_freq_free
IV = 1.080
WOE
200
100
0
-100
-200
<= 0
<= Inf


word_freq_our
IV = 0.853
WOE
200
100
0
-100
-200
<= 0
<= Inf





$`WOE Table for char_freq_xpoint`

	Final.Bin	Total.Count	Total.Distr.	0.Count	1.Count	0.Distr.	1.Distr.	1.Rate	WOE	IV
1	<= 0	1169	50.4%	1015	154	72.2%	16.9%	13.2%	145.5	0.805
2	<= 0.324	570	24.6%	279	291	19.9%	31.9%	51.1%	-47.3	0.057
3	<= Inf	579	25.0%	111	468	7.9%	51.3%	80.8%	-187.0	0.811
5	Total	2318	100.0%	1405	913	100.0%	100.0%	39.4%	NA	1.673

$`WOE Table for word_freq_remove`

	Final.Bin	Total.Count	Total.Distr.	0.Count	1.Count	0.Distr.	1.Distr.	1.Rate	WOE	IV
1	<= 0	1915	82.6%	1387	528	98.7%	57.8%	27.6%	53.5	0.219
2	<= Inf	403	17.4%	18	385	1.3%	42.2%	95.5%	-349.4	1.429
4	Total	2318	100.0%	1405	913	100.0%	100.0%	39.4%	NA	1.647

$`WOE Table for char_freq_dollar`

	Final.Bin	Total.Count	Total.Distr.	0.Count	1.Count	0.Distr.	1.Distr.	1.Rate	WOE	IV
1	<= 0.04675	1738	75.0%	1331	407	94.7%	44.6%	23.4%	75.4	0.378
2	<= Inf	580	25.0%	74	506	5.3%	55.4%	87.2%	-235.4	1.180
4	Total	2318	100.0%	1405	913	100.0%	100.0%	39.4%	NA	1.558

$`WOE Table for word_freq_george`

	Final.Bin	Total.Count	Total.Distr.	0.Count	1.Count	0.Distr.	1.Distr.	1.Rate	WOE	IV
1	<= 0	1925	83.0%	1015	910	72.2%	99.7%	47.3%	-32.2	0.088
2	<= Inf	393	17.0%	390	3	27.8%	0.3%	0.8%	443.6	1.217
4	Total	2318	100.0%	1405	913	100.0%	100.0%	39.4%	NA	1.305

$`WOE Table for word_freq_hp`

	Final.Bin	Total.Count	Total.Distr.	0.Count	1.Count	0.Distr.	1.Distr.	1.Rate	WOE	IV
1	<= 0	1779	76.7%	894	885	63.6%	96.9%	49.7%	-42.1	0.140
2	<= 1.7	308	13.3%	281	27	20.0%	3.0%	8.8%	191.1	0.326

3	<= Inf	231	10.0%	230	1	16.4%	0.1%	0.4%	500.7	0.814
5	Total	2318	100.0%	1405	913	100.0%	100.0%	39.4%	NA	1.280

$`WOE Table for capital_run_length_longest`

	Final.Bin	Total.Count	Total.Distr.	0.Count	1.Count	0.Distr.	1.Distr.	1.Rate	WOE	IV
1	<= 8	747	32.2%	667	80	47.5%	8.8%	10.7%	169.0	0.654
2	<= 55	1120	48.3%	650	470	46.3%	51.5%	42.0%	-10.7	0.006
3	<= Inf	451	19.5%	88	363	6.3%	39.8%	80.5%	-184.8	0.619
5	Total	2318	100.0%	1405	913	100.0%	100.0%	39.4%	NA	1.279

$`WOE Table for word_freq_free`

	Final.Bin	Total.Count	Total.Distr.	0.Count	1.Count	0.Distr.	1.Distr.	1.Rate	WOE	IV
1	<= 0	1687	72.8%	1270	417	90.4%	45.7%	24.7%	68.3	0.305
2	<= Inf	631	27.2%	135	496	9.6%	54.3%	78.6%	-173.2	0.775
4	Total	2318	100.0%	1405	913	100.0%	100.0%	39.4%	NA	1.080

$`WOE Table for word_freq_our`

	Final.Bin	Total.Count	Total.Distr.	0.Count	1.Count	0.Distr.	1.Distr.	1.Rate	WOE	IV
1	<= 0	1423	61.4%	1106	317	78.7%	34.7%	22.3%	81.9	0.360
2	<= Inf	895	38.6%	299	596	21.3%	65.3%	66.6%	-112.1	0.493
4	Total	2318	100.0%	1405	913	100.0%	100.0%	39.4%	NA	0.853

$`WOE Table for capital_run_length_total`

	Final.Bin	Total.Count	Total.Distr.	0.Count	1.Count	0.Distr.	1.Distr.	1.Rate	WOE	IV
1	<= 67	933	40.3%	789	144	56.2%	15.8%	15.4%	127.0	0.513
2	<= Inf	1385	59.7%	616	769	43.8%	84.2%	55.5%	-65.3	0.264
4	Total	2318	100.0%	1405	913	100.0%	100.0%	39.4%	NA	0.777

$`WOE Table for word_freq_edu`

	Final.Bin	Total.Count	Total.Distr.	0.Count	1.Count	0.Distr.	1.Distr.	1.Rate	WOE	IV
1	<= 0	2063	89.0%	1183	880	84.2%	96.4%	42.7%	-13.5	0.016
2	<= Inf	255	11.0%	222	33	15.8%	3.6%	12.9%	147.5	0.180
4	Total	2318	100.0%	1405	913	100.0%	100.0%	39.4%	NA	0.196

Is there a difference?

Running a logistic model with the WOE we get

Deviance Residuals:

Min	1Q	Median	3Q	Max
-5.9628	-0.6218	-0.0010	0.1835	4.8563

Coefficients:	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-1.5621764	0.1110699	-14.065	< 2e-16	***
char_freq_dollar	7.7844205	0.8857794	8.788	< 2e-16	***
char_freq_xpoint	0.3818629	0.1192854	3.201	0.00137	**
word_freq_remove	5.3164782	0.7644838	6.954	3.54e-12	***

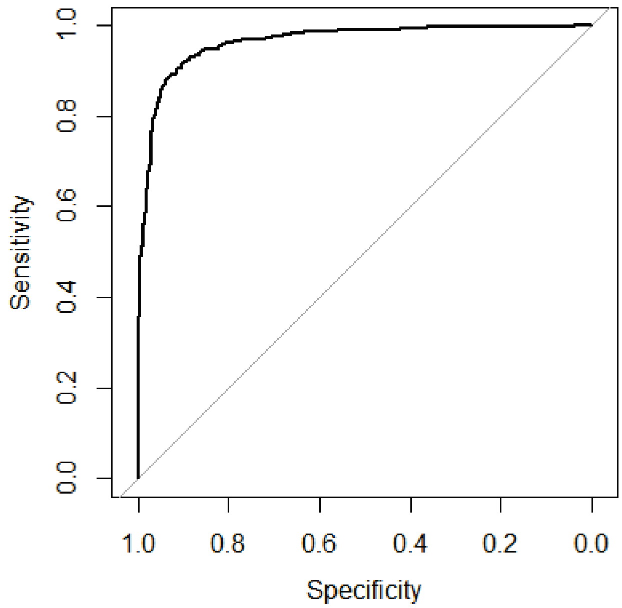
word_freq_hp -2.9688430 0.4116869 -7.211 5.54e-13 ***
capital_run_length_total 0.0004522 0.0001980 2.283 0.02242 *
word_freq_free 1.1855144 0.1751467 6.769 1.30e-11 ***
capital_run_length_longest 0.0162004 0.0026362 6.145 7.98e-10 ***
word_freq_george -6.3246562 1.6246498 -3.893 9.90e-05 ***
word_freq_edu -2.8247355 0.6361340 -4.440 8.98e-06 ***
word_freq_our 0.4961426 0.0950232 5.221 1.78e-07 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)
Null deviance: 3108.2 on 2317 degrees of freedom
Residual deviance: 1268.2 on 2307 degrees of freedom
AIC: 1290.2
Number of Fisher Scoring iterations: 11

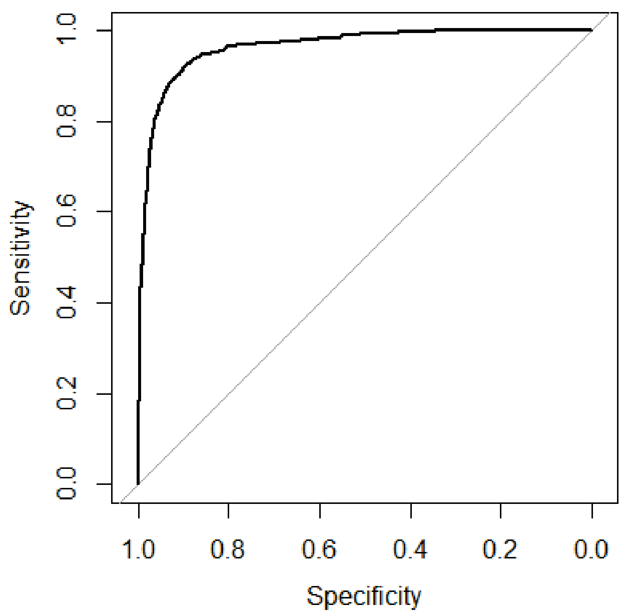
Just spam.df
Deviance Residuals:
Min 1Q Median 3Q Max
-5.8980 -0.5661 0.0000 0.2270 6.1191
Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.486820 0.077288 -19.237 < 2e-16 ***
char_freq_dollar 8.749265 0.668245 13.093 < 2e-16 ***
char_freq_xpoint 0.511925 0.086530 5.916 3.30e-09 ***
word_freq_remove 3.585271 0.383748 9.343 < 2e-16 ***
word_freq_hp -2.752321 0.279845 -9.835 < 2e-16 ***
capital_run_length_total 0.000398 0.000131 3.039 0.00237 **
word_freq_free 1.175765 0.126698 9.280 < 2e-16 ***
capital_run_length_longest 0.015679 0.001727 9.080 < 2e-16 ***
word_freq_george -14.076685 2.255387 -6.241 4.34e-10 ***
word_freq_edu -2.240464 0.302196 -7.414 1.23e-13 ***
word_freq_our 0.437039 0.070883 6.166 7.02e-10 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)
Null deviance: 6170.2 on 4600 degrees of freedom
Residual deviance: 2514.9 on 4590 degrees of freedom
AIC: 2536.9
Number of Fisher Scoring iterations: 12

The AUC for the WOE is 0.9632 and the curve looks like:



The AUC for spam.df is 0.9627



Really don’t see to much difference there.

Next we will see if there is a difference in naïvebayes

Using the WOE date we get:

::: char_freq_dollar (Gaussian)

```
char_freq_dollar         0          1
          mean 0.01177011 0.17638773
          sd   0.07595964 0.37025696
--------------------------------------------------------------------------------
 ::: char_freq_xpoint (Gaussian)
--------------------------------------------------------------------------------
char_freq_xpoint        0         1
          mean 0.1199644 0.5190548
          sd   1.0406681 0.7517288
--------------------------------------------------------------------------------
 ::: word_freq_remove (Gaussian)
--------------------------------------------------------------------------------
word_freq_remove           0           1
          mean 0.006049822 0.269090909
          sd   0.074769239 0.588930552
--------------------------------------------------------------------------------
 ::: word_freq_hp (Gaussian)
--------------------------------------------------------------------------------
word_freq_hp          0          1
      mean 0.86279004 0.01864184
      sd   2.10780868 0.16050741
--------------------------------------------------------------------------------
 ::: capital_run_length_total (Gaussian)
--------------------------------------------------------------------------------
capital_run_length_total        0        1
                mean 165.8121 480.7141
                sd   341.7063 715.5241
--------------------------------------------------------------------------------
 ::: word_freq_free (Gaussian)
--------------------------------------------------------------------------------
word_freq_free          0          1
        mean 0.06018505 0.53208105
        sd   0.32020488 1.04308639
--------------------------------------------------------------------------------
 ::: capital_run_length_longest (Gaussian)
--------------------------------------------------------------------------------
capital_run_length_longest         0         1
                  mean  18.36726 108.01314
                  sd    28.57078 207.59410
--------------------------------------------------------------------------------
 ::: word_freq_george (Gaussian)
--------------------------------------------------------------------------------
word_freq_george           0           1
          mean 1.242925267 0.002070099
          sd   4.209409118 0.045060723
```

::: word_freq_edu (Gaussian)

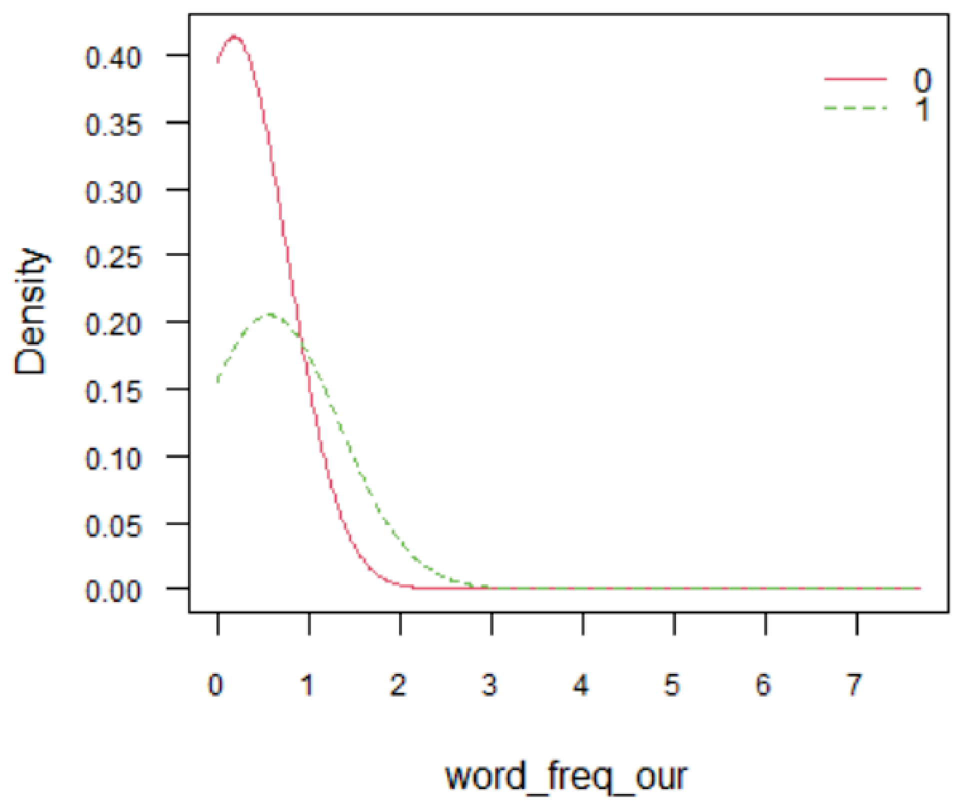
```
word_freq_edu          0          1
        mean 0.29507473 0.01086528
        sd   1.29691907 0.11129332
```

::: word_freq_our (Gaussian)

```
word_freq_our         0         1
        mean 0.1778932 0.5654874
        sd   0.5844510 0.7653683
```

Which also shows
Prior probabilities:

- 0: 0.6061
- 1: 0.3939



Looking at the spam.df data

```
---------------------------------------------------------------------------------
 ::: char_freq_dollar (Gaussian)
---------------------------------------------------------------------------------
char_freq_dollar          0          1
          mean 0.01164849 0.17447821
          sd   0.06964675 0.36047870
---------------------------------------------------------------------------------
 ::: char_freq_xpoint (Gaussian)
---------------------------------------------------------------------------------
char_freq_xpoint         0         1
          mean 0.1099835 0.5137126
          sd   0.8208586 0.7441825
---------------------------------------------------------------------------------
 ::: word_freq_remove (Gaussian)
---------------------------------------------------------------------------------
word_freq_remove          0          1
          mean 0.00938307 0.27540541
          sd   0.11046683 0.57211037
---------------------------------------------------------------------------------
 ::: word_freq_hp (Gaussian)
---------------------------------------------------------------------------------
word_freq_hp          0          1
       mean 0.89547346 0.01747932
       sd   2.07121210 0.16070069
---------------------------------------------------------------------------------
 ::: capital_run_length_total (Gaussian)
---------------------------------------------------------------------------------
capital_run_length_total        0        1
                 mean 161.4709 470.6194
                 sd   355.7384 825.0812
---------------------------------------------------------------------------------
 ::: word_freq_free (Gaussian)
---------------------------------------------------------------------------------
word_freq_free         0         1
         mean 0.0735868 0.5183618
         sd   0.6165739 1.0131699
---------------------------------------------------------------------------------
 ::: capital_run_length_longest (Gaussian)
---------------------------------------------------------------------------------
capital_run_length_longest         0         1
                   mean  18.21449 104.39327
                   sd    39.08479 299.28497
---------------------------------------------------------------------------------
 ::: word_freq_george (Gaussian)
---------------------------------------------------------------------------------
```

word_freq_george 0 1
mean 1.265265423 0.001549917
sd 4.252581229 0.033396282

::: word_freq_edu (Gaussian)

---

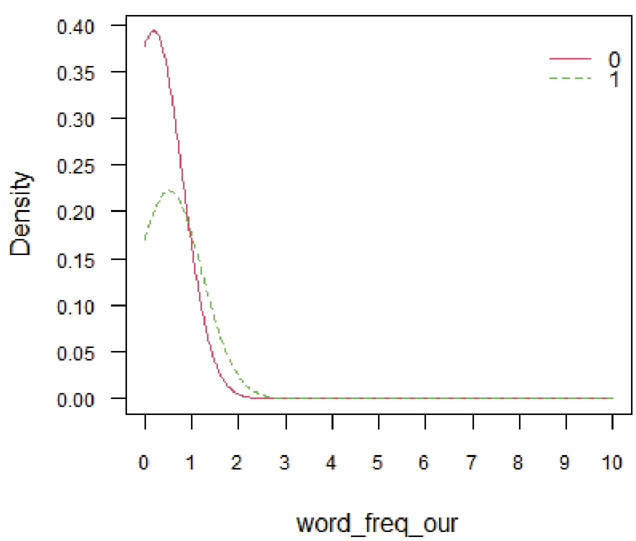
word_freq_edu 0 1
mean 0.28718436 0.01472697
sd 1.15292552 0.13392156

::: word_freq_our (Gaussian)

---

word_freq_our 0 1
mean 0.1810402 0.5139548
sd 0.6145211 0.7071949

- Prior probabilities:
  - 0: 0.606
  - 1: 0.394



Looking at the numbers there is a little bit of a difference but not as much as I thought there would be. However, I would have to say I would gowith WOE model using naïve bayes, since it looks like the most accurate.