

---

## Classifying Text as Song vs. Sonnet

### Description of Dataset:

Our dataset is primarily composed of two different sources. The first part is made up of poems or sonnets from a variety of different poets such as *Shakespeare*, *Keats*, *Frost*, *Shelly*, and *Jackson*, whose texts were all scraped from various webpages. The second part of our dataset is composed of songs from the albums of both *Taylor Swift* and the *Backstreet Boys* which were obtained from a website called *Data.World*, which is an enterprise-level community catalog of open data sets. Each of these components of the entire data set was unfortunately stored in different formats, so a great deal of cleaning had to be done before they could be used.


The sonnets portion was made up entirely of raw text, which had to be parsed into a suitable form. We ended up turning this raw text into a collection of lists of word-strings where each list was composed of a phrase of text with a maximum total length of 12 syllables. To help visualize this, see the example below:

```
['But', 'as', 'the', 'riper', 'should', 'by', 'time', 'decease']
```

This ended up being around 5000 lists of phrases of text long. Unlike the raw text of the sonnets from before, the songs portion were composed of CSV files formatted such that there was a lyrics column and each row represented an entire song from that artist (in the case of the *Taylor Swift* data) or formatted such that each row contained one line from that particular song of that artist (in the case of the *Backstreet Boys*). Both of these were standardized, then parsed and manipulated such that they each were converted into the same form of a list of strings similar to the aforementioned sonnets' format, and eventually came out to be 470 and 1200 phrases long respectively.

Because of the unbalanced sonnet-to-song ratio, we ended up cutting down our sonnet data by 65% percent in a randomized manner. We then did quite a bit of text parsing to get rid of difficult characters, numbers, weird punctuation, or incomplete/empty phrases that were present. After cleaning this data, we finally transformed all of these lists of strings by mapping each string element by syllable to a series of either stressed or unstressed binary categories.

```
1 But [u'B', u'AH1', u'T'] 1
1 as [u'AE1', u'Z'] 1
1 as [u'EH1', u'Z'] 1
0 the [u'DH', u'AH0'] 1
1 the [u'DH', u'AH1'] 1
0 the [u'DH', u'IY0'] 1
riper ---> NOT IN CMU DICT <---
1 should [u'SH', u'UH1', u'D'] 1
1 by [u'B', u'AY1'] 1
1 time [u'T', u'AY1', u'M'] 1
0 decease [u'D', u'IH0', u'S', u'IY1', u'S'] 1
1 decease [u'D', u'IH0', u'S', u'IY1', u'S'] 2
```



```
[1, 1, 1, 2, 1, 1, 1, 2]
```

---

## Statement of Problem:

In this project, we decided to classify whether or not a phrase of text came from a sonnet or a song. To do this we needed to be able to differentiate a sonnet versus a song, and decided upon a feature known as *iambic pentameter*. An *iambic* text is one that is composed of iambs or pairs of *unstressed* then *stressed* syllables. An *iambic pentameter* text is one that is composed of five of these iambs or pairs of unstressed-then-stressed syllables consecutively in a row.

An example of a phrase of text written in *iambic pentameter* is given below:

˘ / ˘ / ˘ / ˘ / ˘ /  
To be | or not | to be | that is | the question

Because it is thought that poems or sonnets are heavily composed of iambic pentameter text compared to other forms of written text like song lyrics, we thought it would be possible to distinguish a sonnet from a song using this concept.

To investigate this classification problem, we decided upon a logistic regression model that would be trained on the aforementioned sonnet and songs dataset. Our response or dependent variable took on a binary value of *\*is sonnet\** or *not*, and our predictor variables that we ran our models on were selected from a grouping of the various *syllabic placements* (eg - syllable 1, syllable 2, syllable 3, etc), as well other interesting variables we were curious about like the *polarity* of the given phrase of text (ie - whether or not a piece of text is positive/negative/neutral) and *subjectivity* of a given phrase of text (ie - text that is phrased in the style of asserting opinions and not statements of fact).

---

## Summary of Methods:

To investigate this classification problem, we built a logistic regression model and analyzed our summary output. We followed this up with a step-wise regression model. For model selection, we found the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) scores for our models, in addition to evaluating our models by running a Wald test to test for the significance of our individual regression coefficients, and a Deviance Chi-squared test to compare our full versus reduced models. We also calculated our pseudo-R<sup>2</sup> values to help evaluate our models and ran a Cross-Validation to assess how our results generalize to an independent data set.

We diagnosed our model by calculating DFFITS in addition to Cook's Distance to find influential points. Further diagnosis was done for multicollinearity issues by calculating our Variance Inflation Factor.

We then created a Confusion Matrix to describe the performance of our classifier model on test data for which we knew the true values. Finally, we plotted an ROC curve to show the relationship between our True Positive Rate versus False Positive Rate at various thresholds.

---

## Explanatory Analysis:

To first explore our data, we created a cross-tabulation of every binary predictor variable that we were considering to use in predicting whether some phrase is a sonnet or not. These predictor variables represent each particular syllabic placement. The cross-tabulation below gives frequency counts for each syllable (s1, s2, s3, ...) for whether it is *stressed*, *unstressed*, or *missing*. The 0s and 1s on the left represent whether or not that phrase is considered a sonnet or not, with 0 being a No and 1 being a Yes.

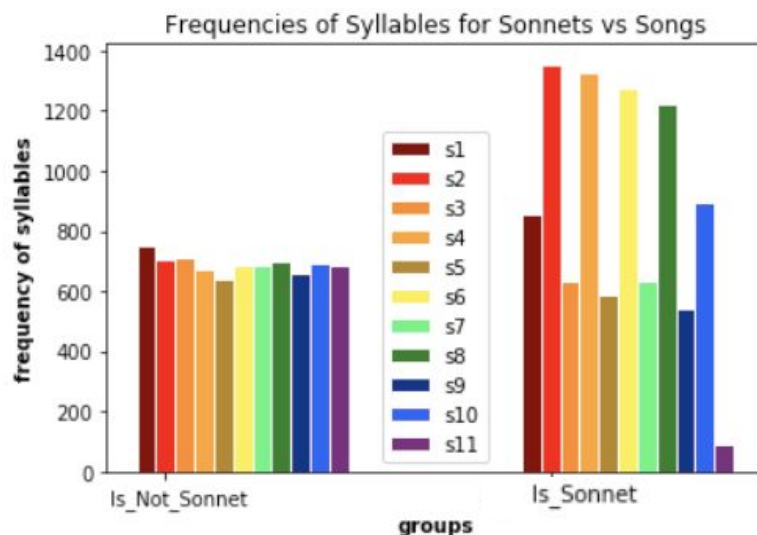
### Cross-Tabulation Analysis

s1	missing	stress	unstress	Total	s7	missing	stress	unstress	Total
sonnet					sonnet				
0	82	1583	40	1705	0	68	1416	221	1705
1	56	830	650	1536	1	115	655	766	1536
Total	138	2413	690	3241	Total	183	2071	987	3241
s2	missing	stress	unstress	Total	s8	missing	stress	unstress	Total
sonnet					sonnet				
0	55	1478	172	1705	0	73	1417	215	1705
1	90	1354	92	1536	1	171	1187	178	1536
Total	145	2832	264	3241	Total	244	2604	393	3241
s3	missing	stress	unstress	Total	s9	missing	stress	unstress	Total
sonnet					sonnet				
0	52	1419	234	1705	0	59	1355	291	1705
1	80	630	826	1536	1	266	540	730	1536
Total	132	2049	1060	3241	Total	325	1895	1021	3241
s4	missing	stress	unstress	Total	s10	missing	stress	unstress	Total
sonnet					sonnet				
0	37	1350	318	1705	0	56	1419	230	1705
1	132	1349	55	1536	1	564	895	77	1536
Total	169	2699	373	3241	Total	620	2314	307	3241
s5	missing	stress	unstress	Total	s11	missing	stress	unstress	Total
sonnet					sonnet				
0	40	1370	295	1705	0	49	1445	211	1705
1	106	573	857	1536	1	1388	90	58	1536
Total	146	1943	1152	3241	Total	1437	1535	269	3241
s6	missing	stress	unstress	Total	s12	missing	stress	unstress	Total
sonnet					sonnet				
0	59	1404	242	1705	0	1705	0	0	1705
1	127	1293	116	1536	1	1518	15	3	1536
Total	186	2697	358	3241	Total	3223	15	3	3241

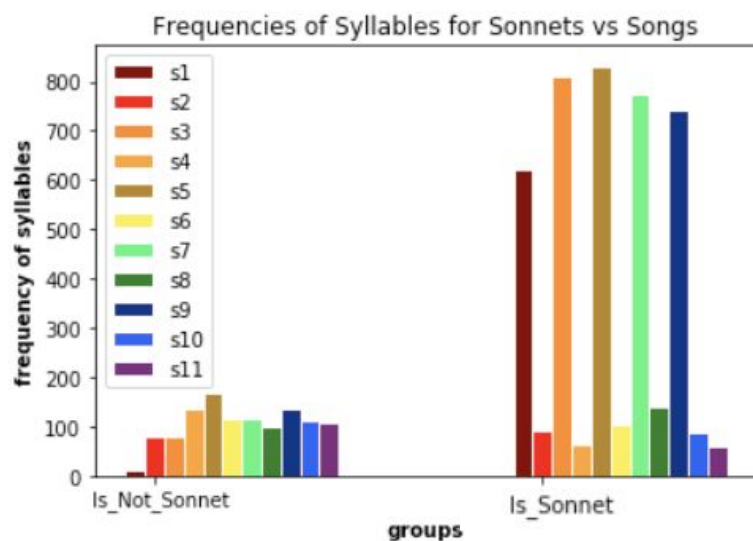
From the above cross-tabulation it was readily apparent that one of the predictor variables, specifically syllable 12, was primarily missing and hence not mapping well to either being stressed or unstressed and so we considered dropping that in our model. The predictor variables of *Polarity* and *Subjectivity* were not included in the above table because of their continuous nature, but we were still open to considering them for further analysis.

To help gain some intuition for the above cross-tabulation we followed up by creating two side-by-side bar plots to help us visualize the frequencies of syllables for each of our predicted categories of sonnets versus songs.

This first side-by-side bar plot gives us the frequencies of each syllable that is *stressed*:



While this second side-by-side bar plot gives us the frequencies of each syllable that is *unstressed*:



As a quick sanity-check, we can notice that when compared to the song group (aka - `is_not_sonnet`), the sonnet group ended up showing a lot more variation in whether or not they exhibited stressed vs. unstressed syllables and actually depicted alternation of unstressed then stressed syllables, which somewhat pre-emptively validates our initial use of this property in distinguishing sonnets versus songs, as songs did not seem to exhibit this alternating unstress-stress syllabic property.

---

## Regression Analysis:

We followed up the above cross-tabulation by creating our basic full logistic regression model with all the binary predictor variables and other predictor variables of interest included:

$$\begin{aligned} \text{sonnet} \sim & C(s1) + C(s2) + C(s3) + C(s4) \\ & + C(s5) + C(s6) + C(s7) + C(s8) \\ & + C(s9) + C(s10) + C(s11) + C(s12) \\ & + \text{syllables} + \text{polarity} + \text{subjectivity} - 1 \end{aligned}$$

After running the logistic regression model we obtained the following summary results:

Generalized Linear Model Regression Results						
Dep. Variable:	sonnet	No. Observations:	3497			
Model:	GLM	Df Residuals:	3469			
Model Family:	Binomial	Df Model:	27			
Link Function:	logit	Scale:	1.0000			
Method:	IRLS	Log-Likelihood:	-337.38			
Date:	Sat, 12 Oct 2019	Deviance:	674.76			
Time:	21:11:14	Pearson chi2:	4.57e+03			
No. Iterations:	22					
Covariance Type:	nonrobust					
	coef	std err	z	P> z	[0.025	0.975]
C(s1) [missing]	27.3439	3.166	8.635	0.000	21.138	33.550
C(s1) [stress]	27.3378	3.111	8.786	0.000	21.239	33.436
C(s1) [unstress]	30.3547	3.179	9.549	0.000	24.124	36.585
C(s2) [T.stress]	-2.3590	0.540	-4.365	0.000	-3.418	-1.300
C(s2) [T.unstress]	-2.6143	0.644	-4.061	0.000	-3.876	-1.352
C(s3) [T.stress]	-0.2662	0.548	-0.486	0.627	-1.340	0.808
C(s3) [T.unstress]	2.1964	0.574	3.823	0.000	1.070	3.322
C(s4) [T.stress]	-1.0454	0.505	-2.072	0.038	-2.034	-0.056
C(s4) [T.unstress]	-2.7177	0.652	-4.170	0.000	-3.995	-1.440
C(s5) [T.stress]	-2.7844	0.511	-5.451	0.000	-3.786	-1.783
C(s5) [T.unstress]	-1.2943	0.529	-2.447	0.014	-2.331	-0.258
C(s6) [T.stress]	-0.6890	0.481	-1.433	0.152	-1.631	0.253
C(s6) [T.unstress]	-0.5344	0.570	-0.937	0.349	-1.652	0.583
C(s7) [T.stress]	-1.1705	0.501	-2.338	0.019	-2.152	-0.189
C(s7) [T.unstress]	0.9806	0.531	1.845	0.065	-0.061	2.022
C(s8) [T.stress]	-0.7422	0.529	-1.403	0.161	-1.779	0.294
C(s8) [T.unstress]	-0.1649	0.606	-0.272	0.785	-1.352	1.022
C(s9) [T.stress]	-2.1207	0.497	-4.271	0.000	-3.094	-1.148
C(s9) [T.unstress]	-1.0279	0.523	-1.965	0.049	-2.053	-0.003
C(s10) [T.stress]	-1.0394	0.389	-2.671	0.008	-1.802	-0.277
C(s10) [T.unstress]	0.3940	0.466	0.846	0.398	-0.519	1.307
C(s11) [T.stress]	-5.3780	0.331	-16.272	0.000	-6.026	-4.730
C(s11) [T.unstress]	-3.6070	0.373	-9.675	0.000	-4.338	-2.876
C(s12) [T.stress]	24.8104	1.57e+04	0.002	0.999	-3.07e+04	3.08e+04
C(s12) [T.unstress]	31.0344	2.93e+04	0.001	0.999	-5.74e+04	5.75e+04
syllables	-1.4967	0.263	-5.696	0.000	-2.012	-0.982
polarity	0.5924	0.381	1.556	0.120	-0.154	1.339
subjectivity	1.0958	0.313	3.499	0.000	0.482	1.710

The first thing we noticed was that syllables 3, 7, and 10, when changed from stressed to unstressed, changed signs from positive coefficients to negative coefficients. This indicates that these variables might be some of the most important when choosing our response variable. The second thing we noticed was that there were one or two variables with very high P-values. So we knew immediately that there needed to be some model selection done in order to improve our prediction. Because we have a large number of predictor variables, after inspecting each of the results associated with the predictor variables we decided to run a Step-Wise Forward Regression in which we added predictor variables of interest based on the test statistics of the estimated coefficients. To do this we started the test with no predictor variables, and at each step we added one more variable where the t-statistic is calculated for the estimated coefficient of each variable not yet in the model and the variable with the highest t-statistic squared was added to our model.

## Model Selection:

We first ran the regular logistic regression on the full model to see how it would perform. Because of our aforementioned P-value problem, we then further implemented a step-wise regression under each of the Akaike Information and Bayesian Information Criteria (AIC and BIC respectively) in order to estimate the quality of each model in our step-wise regression relative to the other models.

Under AIC we ended up with the following predictor variables of interest in our model:

Sonnet ~ C(s11) + C(s1) + C(s3) + C(s5)  
+ C(s7) + C(s9) + syllables + C(s12)  
+ C(s10) + C(s2) + subjectivity + C(s8)  
+ C(s4) + C(s6) + polarity + 1

The following is a summary of the results of our selected model under the AIC criterion:

Generalized Linear Model Regression Results						
=====						
Dep. Variable:	sonnet	No. Observations:	3497			
Model:	GLM	Df Residuals:	3469			
Model Family:	Binomial	Df Model:	27			
Link Function:	logit	Scale:	1.0000			
Method:	IRLS	Log-Likelihood:	-356.37			
Date:	Sat, 12 Oct 2019	Deviance:	712.75			
Time:	16:06:00	Pearson chi2:	5.05e+03			
No. Iterations:	22					
Covariance Type:	nonrobust					
=====						
	coef	std err	z	P> z	[0.025	0.975]
Intercept	32.6798	3.258	10.030	0.000	26.294	39.066
C(s11)[T.stress]	-5.0162	0.309	-16.224	0.000	-5.622	-4.410
C(s11)[T.unstress]	-3.1236	0.348	-8.976	0.000	-3.806	-2.442
C(s1)[T.stress]	-0.7313	0.521	-1.402	0.161	-1.753	0.291
C(s1)[T.unstress]	2.5466	0.582	4.375	0.000	1.406	3.688
C(s3)[T.stress]	-0.7215	0.508	-1.422	0.155	-1.716	0.273
C(s3)[T.unstress]	1.5954	0.533	2.993	0.003	0.551	2.640
C(s5)[T.stress]	-2.3541	0.557	-4.228	0.000	-3.445	-1.263
C(s5)[T.unstress]	-0.9238	0.572	-1.614	0.107	-2.046	0.198
C(s7)[T.stress]	-1.2229	0.478	-2.557	0.011	-2.160	-0.285
C(s7)[T.unstress]	0.6722	0.512	1.312	0.190	-0.332	1.677
C(s9)[T.stress]	-1.3025	0.530	-2.456	0.014	-2.342	-0.263
C(s9)[T.unstress]	0.2016	0.552	0.365	0.715	-0.880	1.283
C(s12)[T.stress]	25.7642	1.46e+04	0.002	0.999	-2.86e+04	2.87e+04
C(s12)[T.unstress]	33.6688	5.13e+04	0.001	0.999	-1e+05	1.01e+05
C(s10)[T.stress]	-0.5658	0.403	-1.404	0.160	-1.356	0.224
C(s10)[T.unstress]	0.9023	0.464	1.946	0.052	-0.006	1.811
C(s2)[T.stress]	-1.6907	0.541	-3.127	0.002	-2.750	-0.631
C(s2)[T.unstress]	-2.3106	0.660	-3.500	0.000	-3.604	-1.017
C(s8)[T.stress]	-1.3422	0.502	-2.673	0.008	-2.326	-0.358
C(s8)[T.unstress]	-0.3479	0.564	-0.616	0.538	-1.454	0.758
C(s4)[T.stress]	-0.7341	0.526	-1.394	0.163	-1.766	0.298
C(s4)[T.unstress]	-2.1002	0.672	-3.125	0.002	-3.417	-0.783
C(s6)[T.stress]	-0.9912	0.454	-2.184	0.029	-1.881	-0.102
C(s6)[T.unstress]	-0.2243	0.521	-0.430	0.667	-1.246	0.797
syllables	-2.0674	0.273	-7.580	0.000	-2.602	-1.533
subjectivity	1.1092	0.293	3.782	0.000	0.534	1.684
polarity	0.5348	0.350	1.526	0.127	-0.152	1.222

While under BIC we ended up with the following predictor variables of interest in our model:

Sonnet ~ C(s11) + C(s1) + C(s3) + C(s5)  
+ C(s7) + C(s9) + syllables + C(s12)  
+ C(s10) + subjectivity + C(s2) + 1

The following is a summary of the results of our selected model under the BIC criterion:

Generalized Linear Model Regression Results						
=====						
Dep. Variable:	sonnet	No. Observations:	3497			
Model:	GLM	Df Residuals:	3476			
Model Family:	Binomial	Df Model:	20			
Link Function:	logit	Scale:	1.0000			
Method:	IRLS	Log-Likelihood:	-376.66			
Date:	Sat, 12 Oct 2019	Deviance:	753.31			
Time:	16:06:06	Pearson chi2:	3.96e+03			
No. Iterations:	22					
Covariance Type:	nonrobust					
=====						
	coef	std err	z	P> z	[0.025	0.975]
Intercept	29.7300	3.136	9.482	0.000	23.584	35.876
C(s11)[T.stress]	-5.0854	0.306	-16.629	0.000	-5.685	-4.486
C(s11)[T.unstress]	-3.2949	0.338	-9.754	0.000	-3.957	-2.633
C(s1)[T.stress]	-0.5052	0.492	-1.028	0.304	-1.469	0.458
C(s1)[T.unstress]	2.8439	0.558	5.097	0.000	1.750	3.937
C(s3)[T.stress]	-0.8412	0.444	-1.893	0.058	-1.712	0.030
C(s3)[T.unstress]	1.5207	0.476	3.194	0.001	0.587	2.454
C(s5)[T.stress]	-2.7894	0.475	-5.870	0.000	-3.721	-1.858
C(s5)[T.unstress]	-1.3613	0.487	-2.796	0.005	-2.315	-0.407
C(s7)[T.stress]	-1.6190	0.399	-4.055	0.000	-2.401	-0.836
C(s7)[T.unstress]	0.2241	0.432	0.519	0.604	-0.622	1.070
C(s9)[T.stress]	-1.6804	0.480	-3.500	0.000	-2.621	-0.740
C(s9)[T.unstress]	-0.2643	0.498	-0.531	0.596	-1.240	0.712
C(s12)[T.stress]	25.3337	1.52e+04	0.002	0.999	-2.97e+04	2.98e+04
C(s12)[T.unstress]	32.9172	5.21e+04	0.001	0.999	-1.02e+05	1.02e+05
C(s10)[T.stress]	-0.5626	0.381	-1.478	0.139	-1.309	0.184
C(s10)[T.unstress]	0.8257	0.442	1.866	0.062	-0.041	1.693
C(s2)[T.stress]	-2.1258	0.471	-4.512	0.000	-3.049	-1.202
C(s2)[T.unstress]	-2.6945	0.604	-4.459	0.000	-3.879	-1.510
syllables	-1.9058	0.267	-7.145	0.000	-2.429	-1.383
subjectivity	1.1113	0.283	3.922	0.000	0.556	1.667

To compare the above three models with our Model 1 being the full logistic regression model, Model A being our final model using AIC, and Model B being our final model after BIC, we decided to run a Cross-Validation. In cross-validation, a model is usually given a data set of known data on which to train (a portion of our dataset set aside for training), and tested against a data of unknown data (a portion of our dataset set aside for testing or validation). Our goal was to test each of our above three models' ability to predict new data that was not used in estimating it, in order to pick the model that best generalizes to an independent data set.

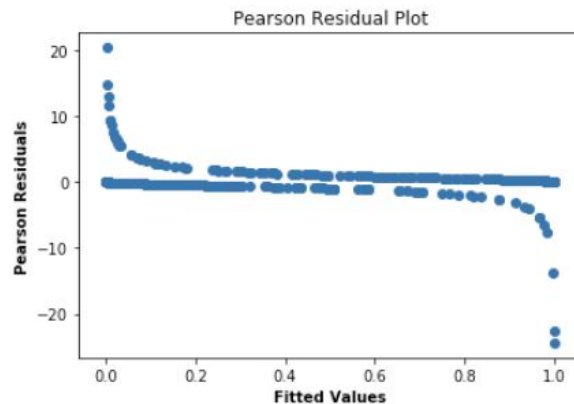
Performing this we can see that our BIC model B performed the best. From our calculations, we can also see that our AIC and BIC scores were low (0.119 and 0.117 respectively), with BIC being the most optimal, leading us to the smaller model with fewer predictors. This smaller model makes sense because BIC penalizes model complexity more heavily.

---

## Model Diagnosis & Evaluation:

To evaluate our model, we first checked the Pseudo- $R^2$  values for a Goodness-of-Fit test. From our summary results table, we report our pseudo- $R^2$  value of 0.8245 which is pretty decent and indicative of possible better model fit.

Afterward, we created a Pearson Residuals plot to check for any initial problems such as non-linearity or skewed normality.



For this plot above, each point represents a phrase of text, where each prediction made by the model is on the x-axis and the accuracy of the model is on the y-axis. The distance from the line at  $y=0$  is how bad the prediction was for that value. Positive values for the residuals mean the prediction was too low, and negative values mean the prediction was too high, with values close to 0 indicating the guess was close to correct. From this, we can see that for most values, our residuals cluster around 0 meaning that those predictions were quite good. However, the errors don't seem to be randomly scattered which potentially indicates that one of the errors might be able to be used to predict another which is problematic (this might be due to autocorrelation which in turn might be due to the time-series nature of text data and unstressed-stressed syllabic patterns). We can also see unequal scatter at the ends of our residual plot indicating some further room for improvement.

To further diagnose our model, we then decided to check for any influential points. To do this we calculated Cook's Distance for both our first model under the AIC criterion (model A) and our second model under the BIC criterion (model B). We calculated Cook's distance and using our rule of thumb of  $D_i > (4/n-p)$  we found influential points at 239 and 253 for models A and B respectively.

Furthermore, we decided to calculate DFFITS for both our aforementioned Model A and Model B, and using the rule of thumb of  $DFFITS \text{ value} > 2 \cdot \sqrt{(p+1)/(n-p-1)}$  we found influential points at 305 and 298 respectively. We kept these values in mind as they could be potential outliers or leverage points.

To further evaluate our model we ended up trying to detect multicollinearity issues by calculating our model's Variance Inflation Factor (VIF) for both our new Model A and new Model B. Multicollinearity exists when two or more predictors in our regression model are moderately or highly correlated with one another.



The following tables give us the VIFs with their corresponding coefficients for both models under consideration:

Model A					Model B				
	VIF	Factor	coefficients	features		VIF	Factor	coefficients	features
0	390.283812		25.467538	Intercept	0	378.451212		24.640724	Intercept
1	2.963959		-5.125366	C(s11)[T.stress]	1	2.801789		-5.012553	C(s11)[T.stress]
2	1.573601		-3.307410	C(s11)[T.unstress]	2	1.534998		-3.362928	C(s11)[T.unstress]
3	5.619393		-0.567403	C(s1)[T.stress]	3	5.619551		-0.474076	C(s1)[T.stress]
4	5.855093		2.544954	C(s1)[T.unstress]	4	5.837783		2.694535	C(s1)[T.unstress]
5	5.981673		-1.196201	C(s7)[T.stress]	5	7.968339		-0.678822	C(s3)[T.stress]
6	6.011415		0.724096	C(s7)[T.unstress]	6	7.946578		1.546412	C(s3)[T.unstress]
7	7.968535		-0.613686	C(s3)[T.stress]	7	7.370275		-2.255391	C(s5)[T.stress]
8	7.949118		1.617261	C(s3)[T.unstress]	8	7.334554		-0.829101	C(s5)[T.unstress]
9	7.369812		-2.263573	C(s5)[T.stress]	9	5.973563		-1.045838	C(s7)[T.stress]
10	7.332100		-0.849826	C(s5)[T.unstress]	10	6.006734		0.776879	C(s7)[T.unstress]
11	4.242202		-1.124240	C(s9)[T.stress]	11	3.840013		-1.125800	C(s9)[T.stress]
12	4.282638		0.330061	C(s9)[T.unstress]	12	3.651769		0.163042	C(s9)[T.unstress]
13	2.276174		-0.566918	C(s10)[T.stress]	13	3.294270		-0.927640	C(s6)[T.stress]
14	1.865061		0.832463	C(s10)[T.unstress]	14	3.096931		-0.247185	C(s6)[T.unstress]
15	3.457331		-1.676572	C(s2)[T.stress]	15	3.444428		-1.756127	C(s2)[T.stress]
16	3.245905		-1.720490	C(s2)[T.unstress]	16	3.237476		-1.730413	C(s2)[T.unstress]
17	3.702222		-0.873735	C(s4)[T.stress]	17	3.242836		-1.331065	C(s8)[T.stress]
18	3.518350		-2.066392	C(s4)[T.unstress]	18	2.926527		-0.512220	C(s8)[T.unstress]
19	3.250811		-1.421777	C(s8)[T.stress]	19	3.700727		-0.833969	C(s4)[T.stress]
20	2.938853		-0.634043	C(s8)[T.unstress]	20	3.516839		-1.763912	C(s4)[T.unstress]
21	3.293824		-0.954309	C(s6)[T.stress]	21	2.089397		-1.362752	syllables
22	3.094335		-0.372937	C(s6)[T.unstress]	22	1.011150		0.635996	polarity
23	2.207825		-1.414603	syllables					
24	1.019263		1.142091	subjectivity					

A VIF of 1 means that our associated predictor variable is not correlated with other variables, with higher values indicative of greater correlation with other variables. Since values around 4 are considered moderate to high and values greater than 10 are usually considered very high, we can see that we unfortunately do have some multicollinearity issues in both of our models, which makes sense given that many of the phrases in our data set consist of alternating stressed and unstressed syllables. It seems that our predictor variables of the syllables s3 through s10 are something we should be cognizant of. Multicollinearity introduces problems in our models such as inflating the estimated standard error of coefficients, making our t-statistics smaller and p-values larger which makes it more difficult to have rejections.

With the aforementioned in mind, to get an idea of the significance of individual regression coefficients we ran a Wald Test to test for significance of individual regression coefficients, with our null hypothesis  $H_0: B_i = 0$  versus our alternative hypothesis  $H_a: B_i \neq 0$ , with a rejection rule  $Z_s > Z_{\alpha/2}$ . We found our significant predictors to be syllables s1, s2, s4, s11, and subjectivity.

Finally, to evaluate our model we ran a Deviance Chi-Squared test to compare our reduced and full models. We calculated the deviance of our full model to be 674.76, and our reduced model to be 684.53. We then calculated delta- $G^2$  to be 9.76 and compared this with our Chi-Squared value of 40.11. Since our delta- $G^2$  value is smaller than our Chi-squared, we failed to reject our null hypothesis in favor of the alternative which is our reduced model.

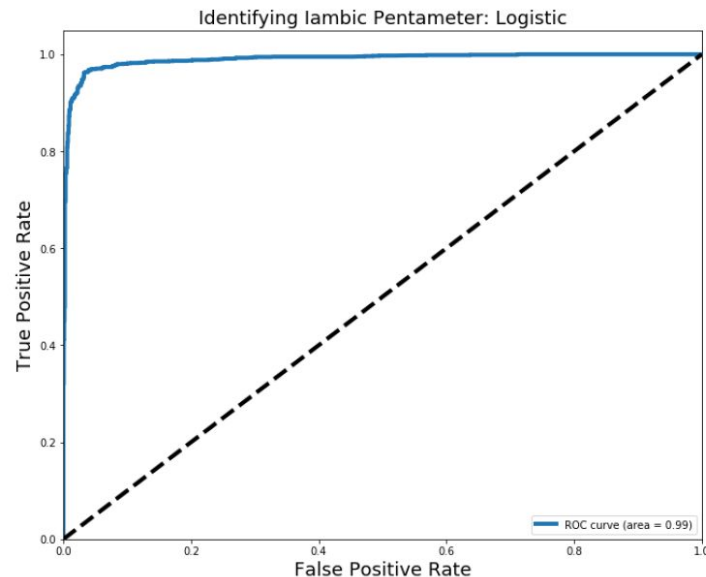
## Final Model Choice & Interpretation:

After the aforementioned model evaluation, we ended up choosing Model B which was one of our reduced models chosen with Step-Wise Regression underneath the Bayesian Information Criterion. It was also the least complex out of all the models above with the fewest number of predictor variables. To estimate the performance of our classifier model on test data for which we knew the true values, we ended up creating a Confusion Matrix:

Predicted	0.0	1.0	All
True			
0.0	1650	55	1705
1.0	74	1718	1792
All	1724	1773	3497

This matrix plots Predicted Values on the x-axis with Actual values on the y-axis. This can be quite useful for measuring Recall, Precision, Specificity, and Accuracy of our model if needed. From the above table, we can see that it gives us our True positives in the bottom right corner, False positives in the top right corner, True negatives in the top left corner, and False negatives in the bottom left corner. The values we received were quite good in that what we predicted was a sonnet actually turned out to be a sonnet (bottom right), and what we predicted was not a sonnet (a song) actually turned out to be a song (top left), with very little Type I (false positives), and Type II (false negatives) errors.

Finally, we created a ROC curve which gives us the plot of the true positive rate against the false-positive rate for the different cut points of our test:



This ROC curve demonstrates the tradeoff between Sensitivity (True positive / (True Positive + False Negative)) and Specificity (True Negative / (True Negative + False Positive)), where any increase in sensitivity will be accompanied by a decrease in specificity. Since our curve closely follows the upper left-hand border of the top portion of our ROC space and is relatively far away from the 45-degree placed line  $y=x$  (where a 50:50 guess would happen to be), we have a more accurate test.

---

## Summary:

In summary, we started with a data set composed of the text of both poems/sonnets and songs. We wanted to classify whether or not a given piece of text was a sonnet or a song. To do this we leveraged a concept known as iambic pentameter which reflects the pattern of unstressed-then-stressed syllables in an alternating consecutive manner. Because sonnets are thought to be often written in this stylistic manner compared to other forms of written word, we decided to use this concept to differentiate between our sonnets versus songs. To do this, we first mapped our text data to syllables marked by the binary category of stressed versus unstressed. These binary categories for each of the 12 syllables used were combined with other variables of interest like subjectivity and polarity of a piece of phrase of text, and in total made up our predictor variables to be fed into our initial full model.

Because this was a binary classification problem, we ended up using a Logistic Regression Model. To improve our model, we narrowed down our predictor variables by performing Step-Wise Regression under both the AIC and BIC criterions yielding two new models for comparison. We evaluated our models through a series of steps, first diagnosing Goodness-of-Fit and creating Residual Plots. We then found possible Influential Points and investigated Multicollinearity issues that were present among our series of stressed and unstressed syllables, from which we then calculated Variance Inflation factors. We then ran a Wald Test and Deviance-Chi-Squared Test to evaluate our full versus reduced models.

We finally decided upon Model B selected under BIC which was the least complex out of all our models. We further checked the performance of our model by creating a Confusion Matrix and then plotting an ROC curve to check for accuracy which turned out to be quite good.

Further potential avenues for improvement could be made in choosing our model. One way is to remove some of our variables that have higher colinear relationships. Secondly, we could use interaction terms in our model for each iambic foot which could significantly reduce the multicollinearity between our variables. Generally speaking, we could add more data to our data set especially the number of songs. Other than those few improvements, we were happy with how our model performed.

---

## Appendix:

(all the following code & data used can be found in a GitHub repository here: [https://github.com/amadorschulze92/iambic\\_songs](https://github.com/amadorschulze92/iambic_songs))

### I. Data Cleaning Code:

See provided *ingest\_text.ipynb* or *ingest\_text.py*  
and *data\_cleaning.ipynb* or *functions.py*

### II. Data Exploration & Plotting Code:

See provided *my\_graphs.ipynb*

### III. Data Analysis Code:

See provided *model.ipynb*