Predicting Gender with Email Body Text: Binary Classification Supervised Learning with The Enron Email Dataset

James Bush, Springboard Capstone I

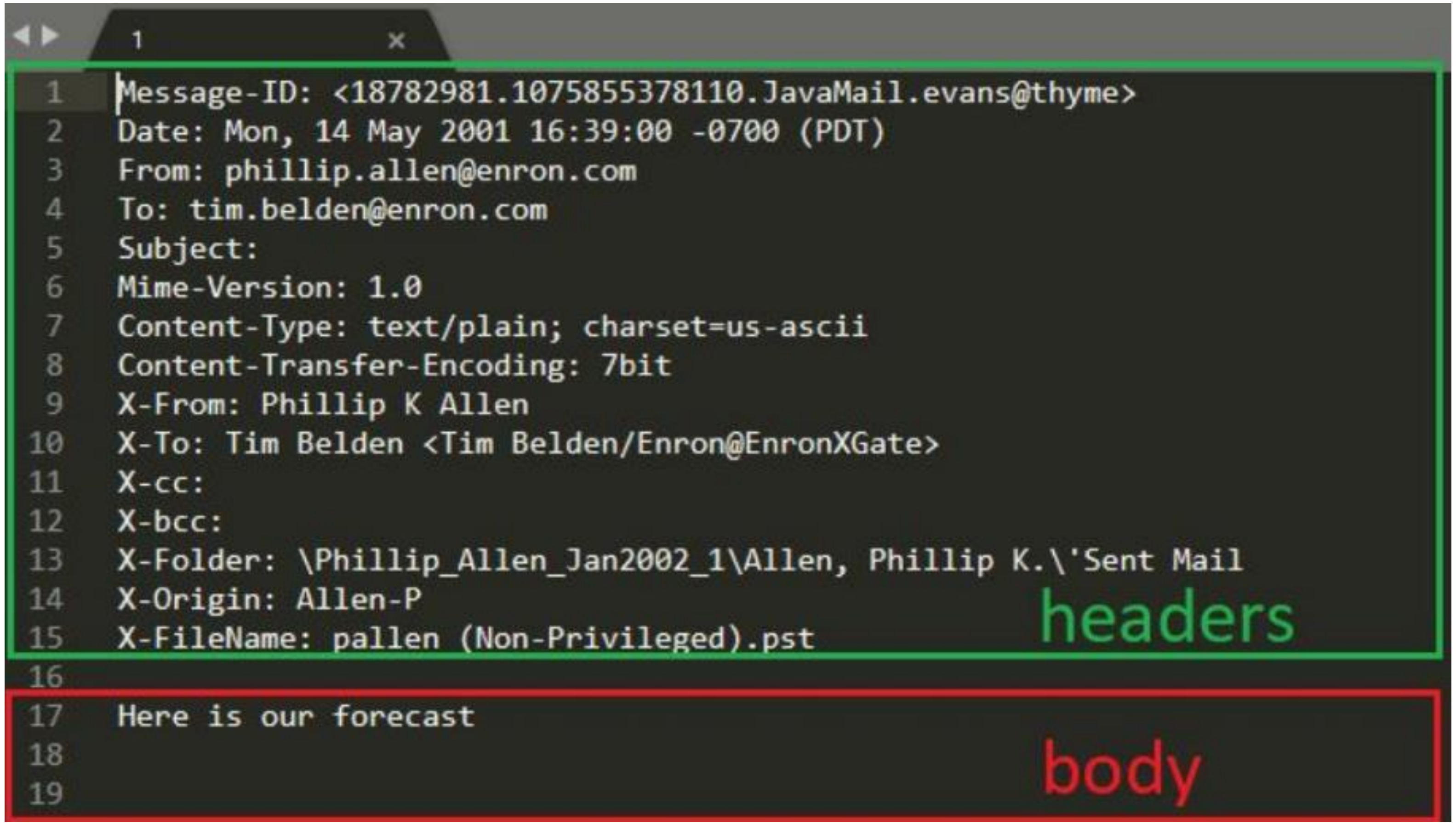
TABLE OF CONTENTS

DEFINING THE PROBLEM

- Can **text** contained in an email be used to **predict** the sender's **gender?**
- Feature Variables, Target Variables
- Genuine vs. Non-Genuine Text

DATA COLLECTION

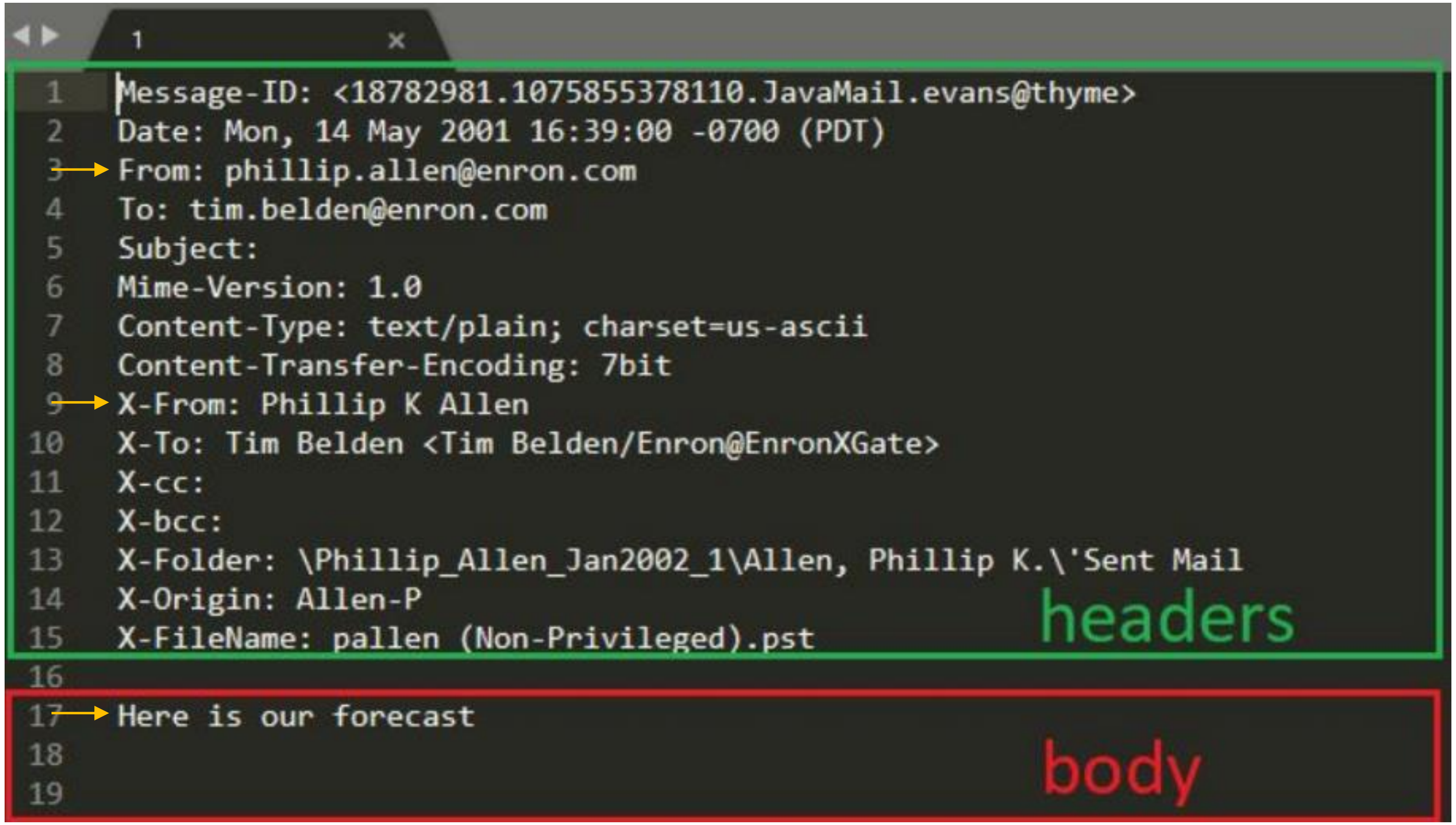
- Enron Email Dataset from Carnegie Mellon University School of Computer Science
- Example:



```
Message-ID: <18782981.1075855378110.JavaMail.evans@thyme>
Date: Mon, 14 May 2001 16:39:00 -0700 (PDT)
From: phillip.allen@enron.com
To: tim.belden@enron.com
Subject:
Mime-Version: 1.0
Content-Type: text/plain; charset=us-ascii
Content-Transfer-Encoding: 7bit
X-From: Phillip K Allen
X-To: Tim Belden <Tim Belden/Enron@EnronXGate>
X-cc:
X-bcc:
X-Folder: \Phillip_Allen_Jan2002_1\Allen, Phillip K.\'Sent Mail
X-Origin: Allen-P
X-FileName: pallen (Non-Privileged).pst

Here is our forecast
```

DATA COLLECTION

- Variables of interest

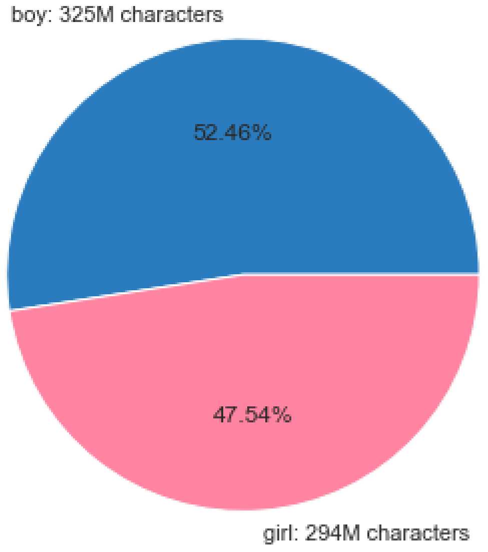


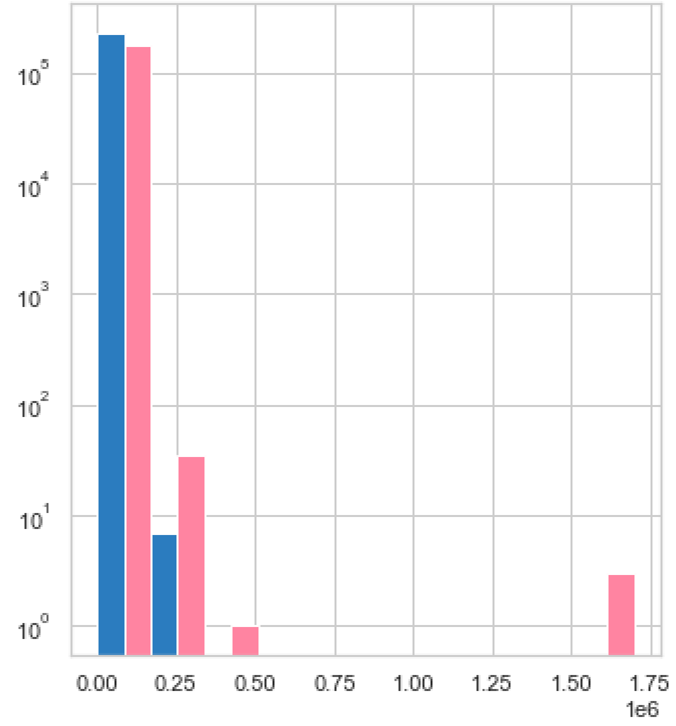
```
Message-ID: <18782981.1075855378110.JavaMail.evans@thyme>
Date: Mon, 14 May 2001 16:39:00 -0700 (PDT)
From: phillip.allen@enron.com
To: tim.belden@enron.com
Subject:
Mime-Version: 1.0
Content-Type: text/plain; charset=us-ascii
Content-Transfer-Encoding: 7bit
X-From: Phillip K Allen
X-To: Tim Belden <Tim Belden/Enron@EnronXGate>
X-cc:
X-bcc:
X-Folder: \Phillip_Allen_Jan2002_1\Allen, Phillip K.\'Sent Mail
X-Origin: Allen-P
X-FileName: pallen (Non-Privileged).pst

Here is our forecast
```

DATA WRANGLING

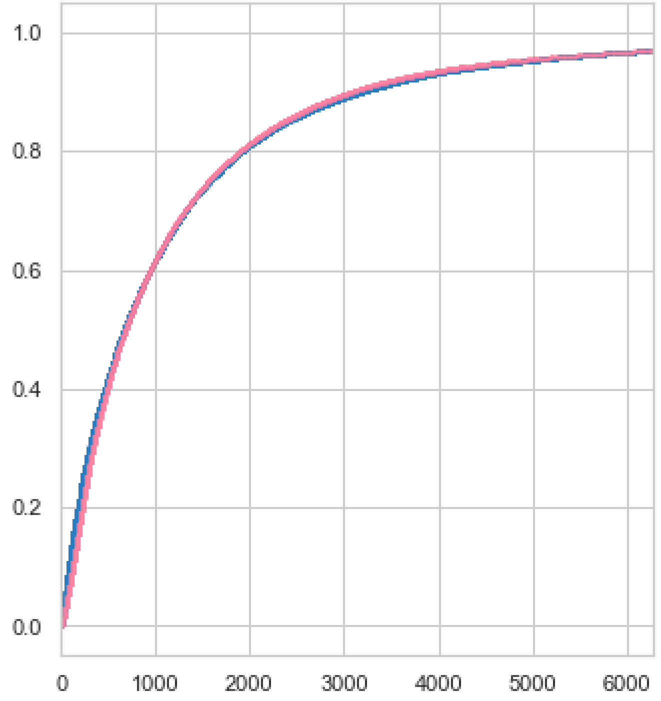
- **Collecting Gender**
- **Data Cleaning**
- **Cosine Similarity**
- **Preprocessing**

Initial Data Picture





	boy	girl
count	235804.00	174119.00
mean	1408.53	1693.43
std	3325.31	9332.59
min	1.00	1.00
25%	231.00	279.00
50%	670.00	684.00
75%	1585.00	1560.00
max	248665.00	1697165.00



DATA WRANGLING

Collecting Gender

Separate, Clean First Name for Scraping

```
cd['gender_query'] = cd['clean_name'].str.extract('^([A-Za-z\'-]+) [A-Za-z\'-]+$')
cd.head()
```

	m_from	m_from_cleaned	x_from	x_from_cleaned	clean_name	gender_query
0	phillip.allen@enron.com	phillip allen	phillip k allen	phillip allen	phillip allen	phillip
1	ina.rangel@enron.com	ina rangel	ina rangel	ina rangel	ina rangel	ina
2	1.11913372.-2@multexinvestornetwork.com		multex investor <1.11913372.-2@multexinvestorn...	multex investor	multex investor	multex
7	rebecca.cantrell@enron.com	rebecca cantrell	rebecca w cantrell	rebecca cantrell	rebecca cantrell	rebecca
9	paul.kaufman@enron.com	paul kaufman	paul kaufman	paul kaufman	paul kaufman	paul

DATA WRANGLING

Data Cleaning

Filter Out Emails Based On:
- Drop duplicates
- Drop NaN values
- ‘@enron’ domain name
- ‘Copyright’ string
- Automated output reports

Remove Strings:
- URLs
- Email Addresses
- [word] colon (i.e. *From:*)

Split Emails Based On:
- Consecutive *m-dash* characters (‘-‘)
- Patterns preceding *forwarded text*
- Patterns preceding *reply text*
- Timestamps associated with above

DATA WRANGLING

Cosine Similarity

```
[INDEX 1058]
We should definitely bring him in for an interview.

From:  Lexi Elliott                          02/09/2001 09:04 AM

To: Richard Causey/Corp/Enron@ENRON
cc: Mark E Lindsey/GPGFIN/Enron@ENRON, Mike Deville/HOU/ECT@ECT, Sally
Beck/HOU/ECT@ECT

Subject: Summer Internship

This candidate is currently working in Houston.  Since our schedules are full
on-campus, it would be very easy to bring him in-house for interviews.
Please let me know if you are interested.

Thank you,

Lexi

----------------------
```

```
[INDEX 1121]
Please give me Amy's e-mail address.  I'd like to get in touch with her.
Thanks.

Emily Sellers
02/06/2001 01:16 PM

To: Steven J Kean/NA/Enron@Enron
cc:
Subject: follow up to interview

Steve, Amy Kim asked me to forward this to you.  She wasn't sure if she had
the correct email address.

Emily Sellers
----------------------
```

DATA WRANGLING

Preprocessing

- Remove non-word characters
- Remove underscore character
- Remove single characters
- Remove numbers
- Reduce space character multiples
- Remove stop words
- **Remove names included in the gender-name key**
- Lemmatize words

DATA WRANGLING

ORIGINAL BODY

```
We should definitely bring him in for an interview.
```

SPLIT ON

```
        From:  Lexi Elliott                          02/09/2001 09:04 AM
```

```
To: Richard Causey/Corp/Enron@ENRON
cc: Mark E Lindsey/GPGFIN/Enron@ENRON, Mike Deville/HOU/ECT@ECT, Sally
Beck/HOU/ECT@ECT

Subject: Summer Internship

This candidate is currently working in Houston.  Since our schedules are full
on-campus, it would be very easy to bring him in-house for interviews.
Please let me know if you are interested.

Thank you,

Lexi

---------------------- Forwarded by Lexi Elliott/NA/Enron on 02/09/2001 09:09
AM ---------------------------


Judd Eisenberg <judde@mail.utexas.edu> on 02/06/2001 01:11:05 AM
To: lexi.elliott@enron.com
cc:

Subject: Summer Internship


Lexi Elliot,

Hi, my name is Judd Eisenberg, and I am a business student at the
University of Texas who is seeking the summer analyst internship at Enron.
I received an opportunity to meet you at a reception dinner that Enron
```

POST DATA CLEANING

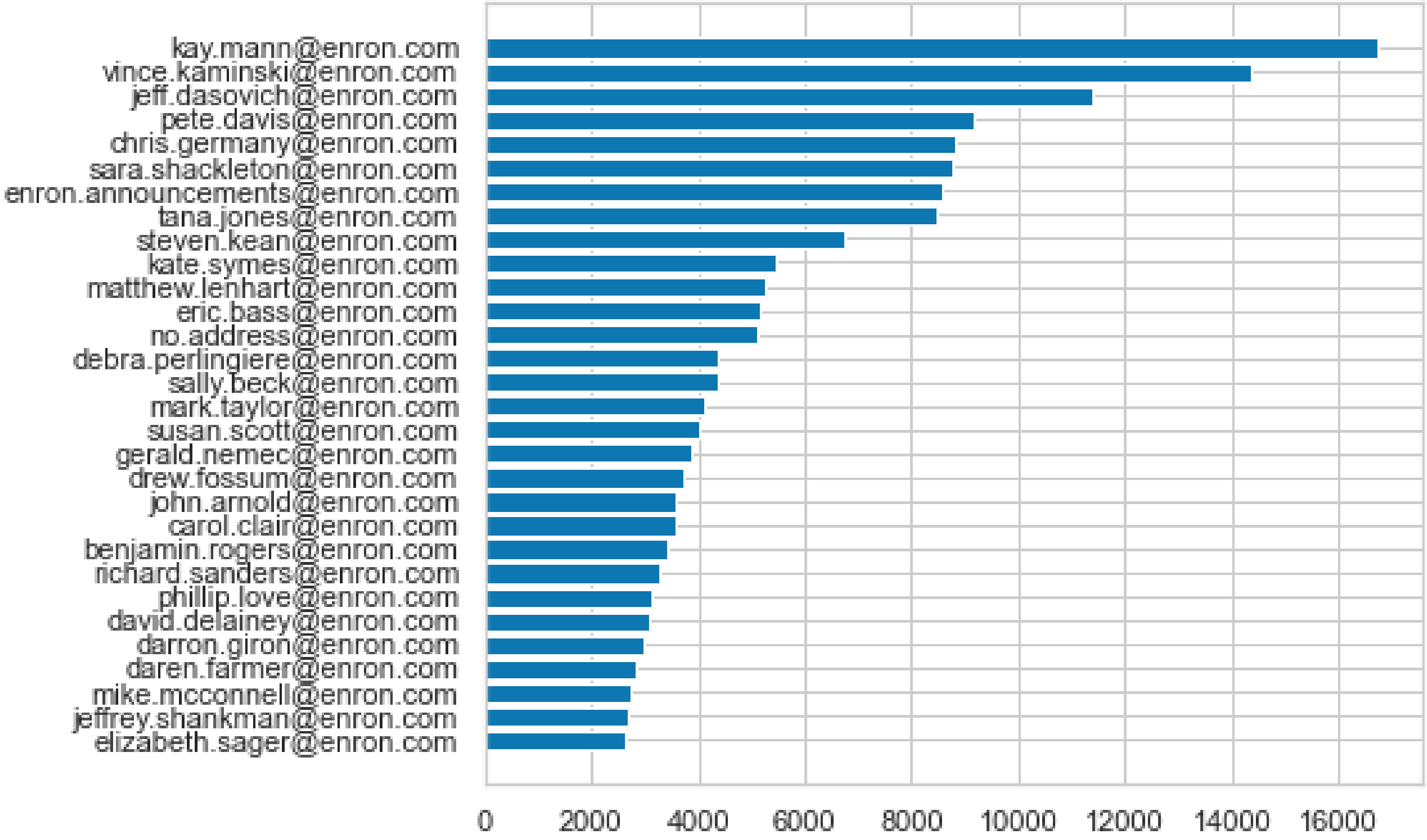
```
We should definitely bring him in for an interview.
```

POST PREPROCESSING

```
definitely bring interview
```

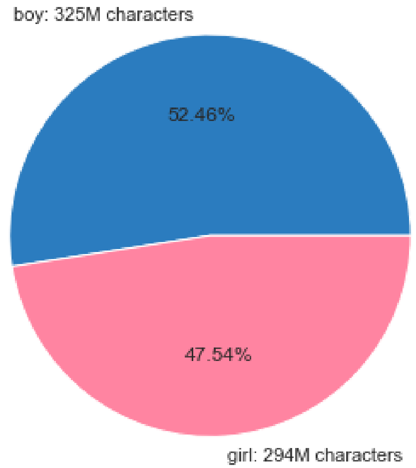
EXPLORATORY DATA ANALYSIS

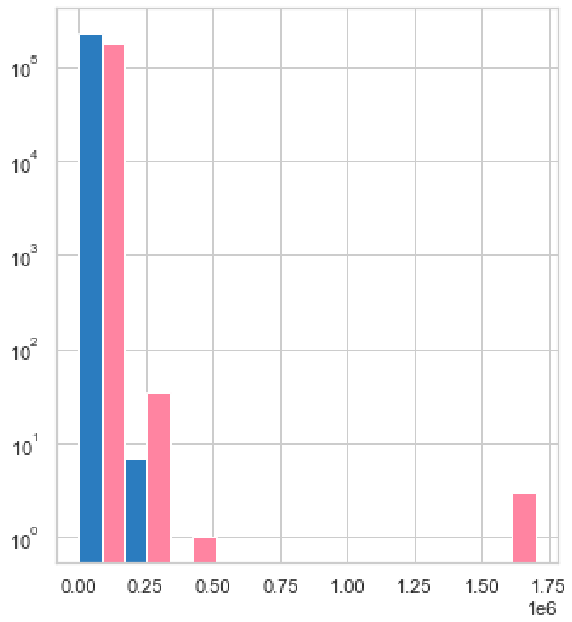
Number of Emails by Sender



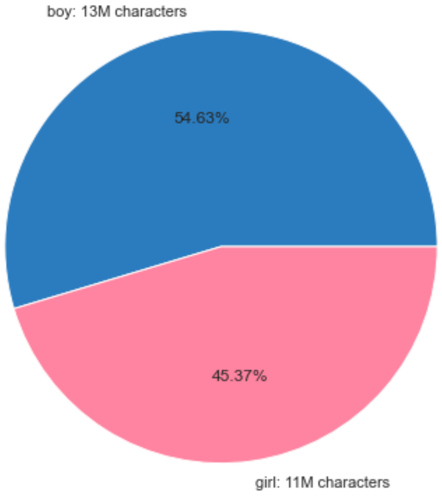
EXPLORATORY DATA ANALYSIS

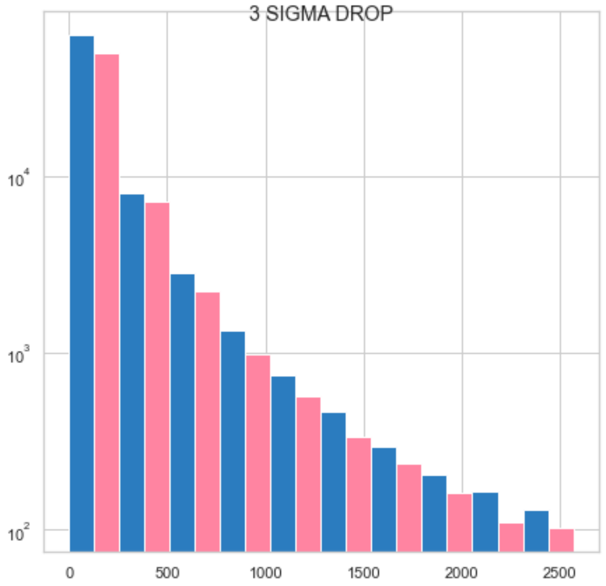
Initial Data Picture





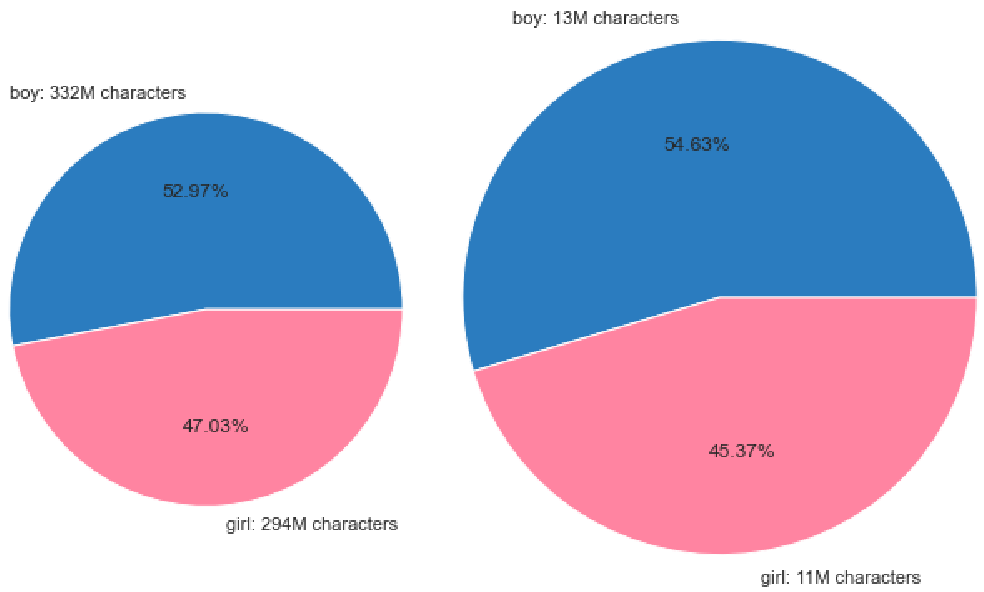
Pre-Modeling Data Picture



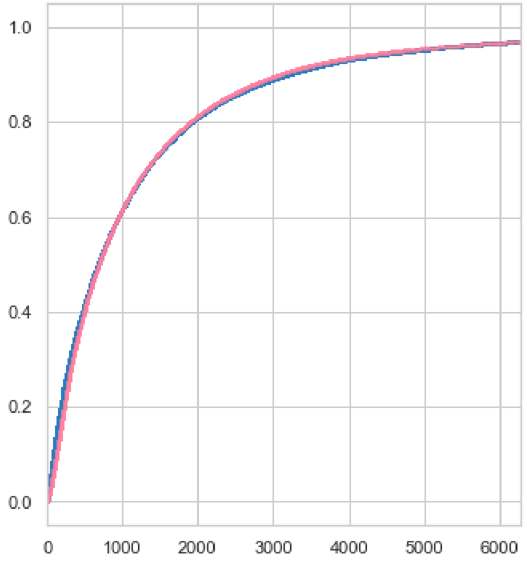


EXPLORATORY DATA ANALYSIS

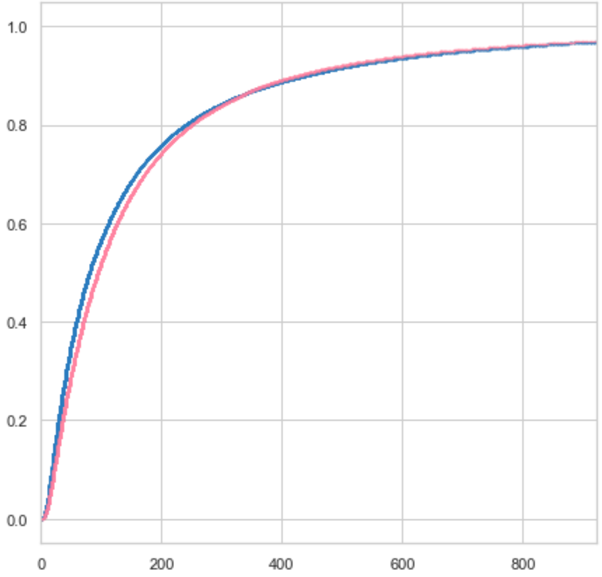
Character Count Change
From Initial to Pre-Modeling



	boy	girl
count	235804.00	174119.00
mean	1408.53	1693.43
std	3325.31	9332.59
min	1.00	1.00
25%	231.00	279.00
50%	670.00	684.00
75%	1585.00	1560.00
max	248665.00	1697165.00



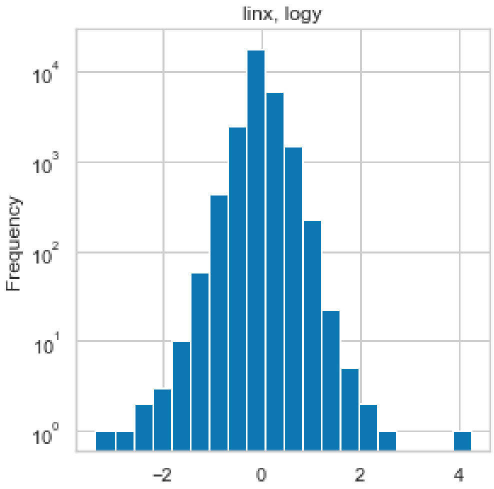
	boy	girl
count	76597.00	61665.00
mean	180.93	186.65
std	289.02	279.92
min	1.00	1.00
25%	35.00	43.00
50%	80.00	94.00
75%	192.00	205.00
max	2565.00	2557.00

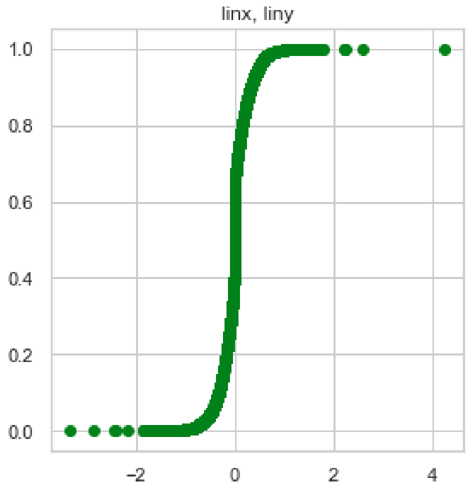


FEATURE ANALYSIS

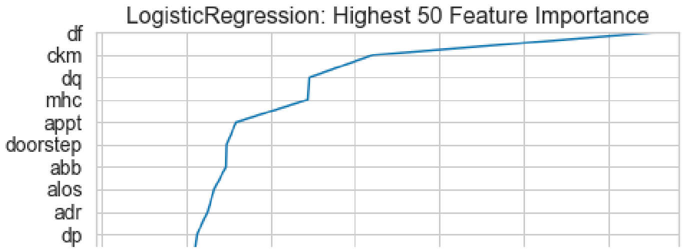
Evaluating Thresholds – No Restrictions

Feature Coefficients Histogram, ECDF





Feature Returns



Feature Returns with Data Points

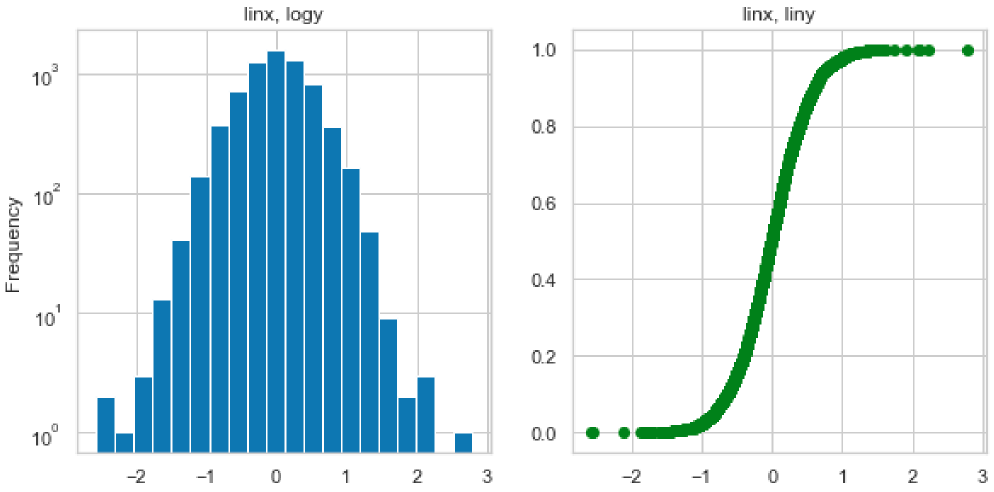
feature_name	feature_coef	feature_frequency	character_count
pl	-3.365496	259	2
bt	-2.879711	100	2
jmf	-2.445996	67	3
dg	-2.403187	145	2
mat	-2.180302	68	3
cgy	-1.882465	23	3
executables	-1.857019	15	11
kk	-1.833449	25	2
kh	-1.806247	29	2
db	-1.762353	32	2

feature_name	feature_coef	feature_frequency	character_count
df	4.257436	179	2
ckm	2.606287	47	3
dq	2.229836	20	2
mhc	2.219864	30	3
appt	1.797386	24	4
doorstep	1.741485	39	8
abb	1.737722	40	3
alos	1.667401	22	4
adr	1.629356	34	3
dp	1.566950	59	2

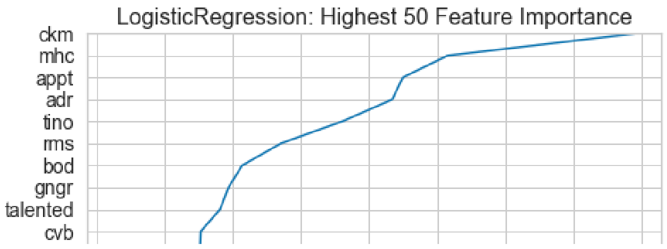
FEATURE ANALYSIS

Evaluating Thresholds – First Adjustment
Word Count: 7df - 60% df
Character Count: 3, 13

Feature Coefficients Histogram, ECDF



Feature Returns



Feature Returns with Data Points

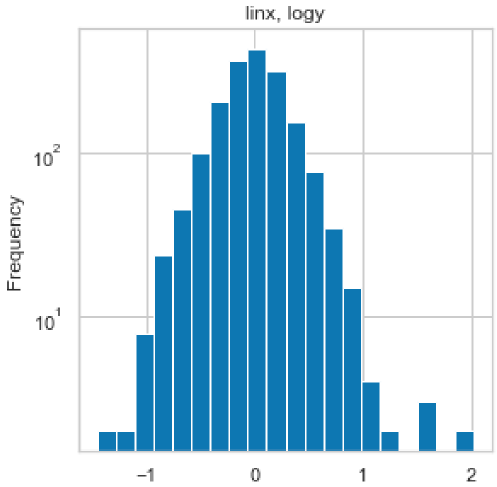
feature_name	feature_coef	feature_frequency	character_count
mat	-2.560905	71	3
jmf	-2.542075	77	3
cgas	-2.123214	59	4
executables	-1.874143	16	11
gtv	-1.825940	23	3
shout	-1.799436	35	5
cng	-1.750343	73	3
hgm	-1.724560	11	3
latter	-1.719614	35	6
aec	-1.713489	14	3

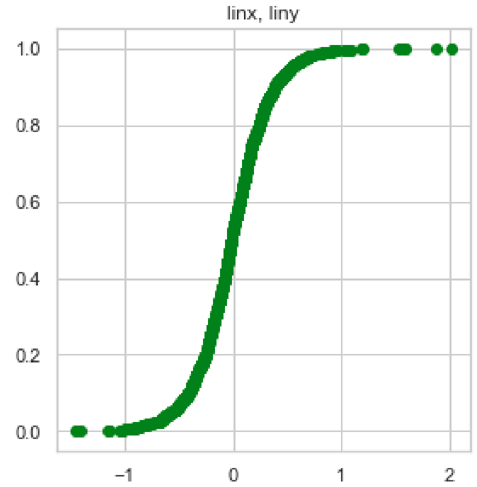
feature_name	feature_coef	feature_frequency	character_count
ckm	2.782829	54	3
mhc	2.231569	35	3
appt	2.099367	26	4
adr	2.068282	43	3
tino	1.921104	31	4
rms	1.741359	19	3
bod	1.627470	18	3
gngr	1.587551	25	4
talented	1.562603	13	8
cvb	1.506041	8	3

FEATURE ANALYSIS

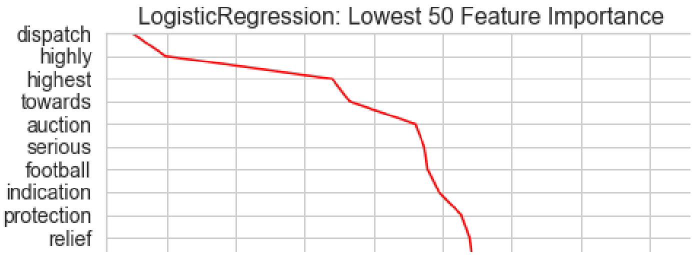
Evaluating Thresholds – Second Adjustment
Word Count: 50df - 50% df
Character Count: 5, 11

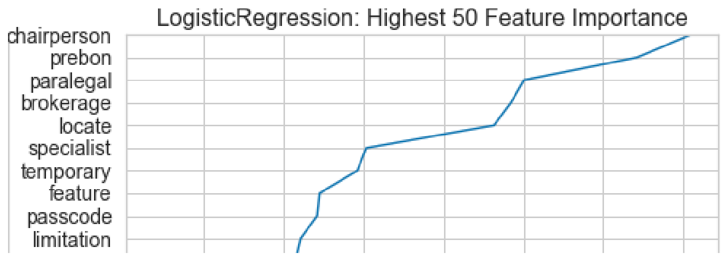
Feature Coefficients Histogram, ECDF





Feature Returns



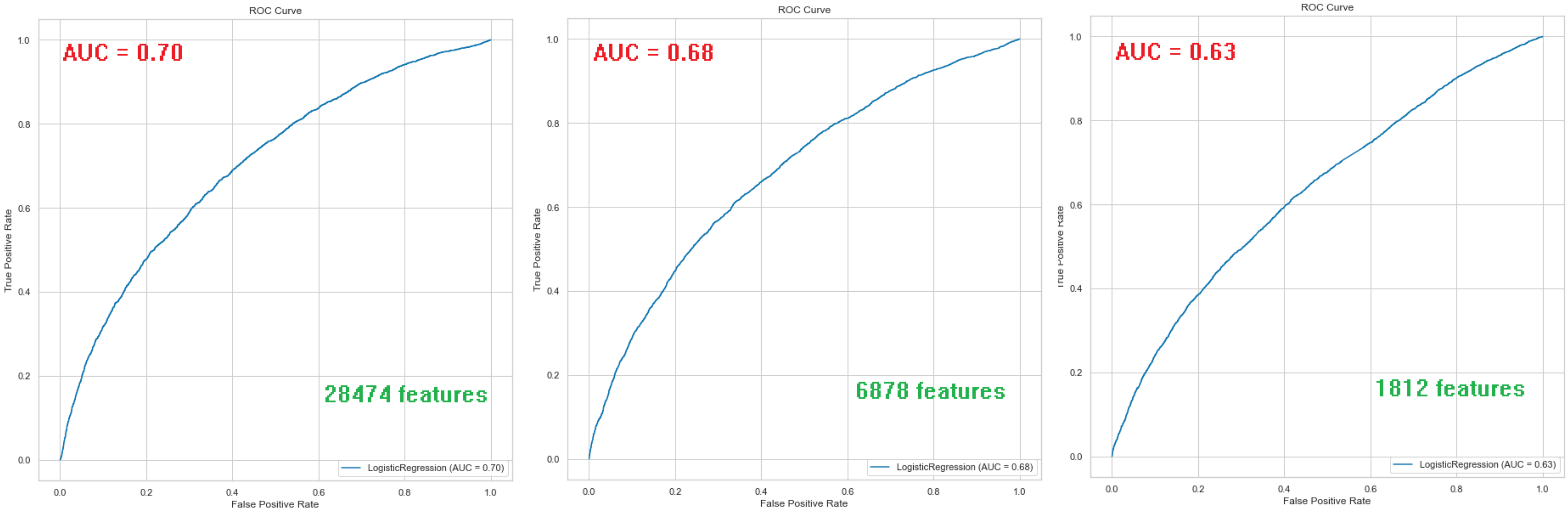


Feature Returns with Data Points

feature_name	feature_coef	feature_frequency	character_count
dispatch	-1.448862	68	8
highly	-1.401186	94	6
highest	-1.160044	59	7
towards	-1.134885	83	7
auction	-1.039440	140	7
serious	-1.026771	68	7
football	-1.021749	77	8
indication	-1.005445	55	10
protection	-0.973110	63	10
relief	-0.961283	71	6

feature_name	feature_coef	feature_frequency	character_count
chairperson	2.018544	82	11
prebon	1.884477	115	6
paralegal	1.601634	65	9
brokerage	1.568712	87	9
locate	1.526571	62	6
specialist	1.207088	201	10
temporary	1.184306	79	9
feature	1.089935	80	7
passcode	1.083886	80	8
limitation	1.042528	78	10

FEATURE ANALYSIS



Thresholds for Model:

- **Feature Frequency: Min df: 7, Max df: 70%**
- **Feature Characters: Min chars: 4, Max chars: 13**

MODEL IMPLEMENTATION, RESULTS

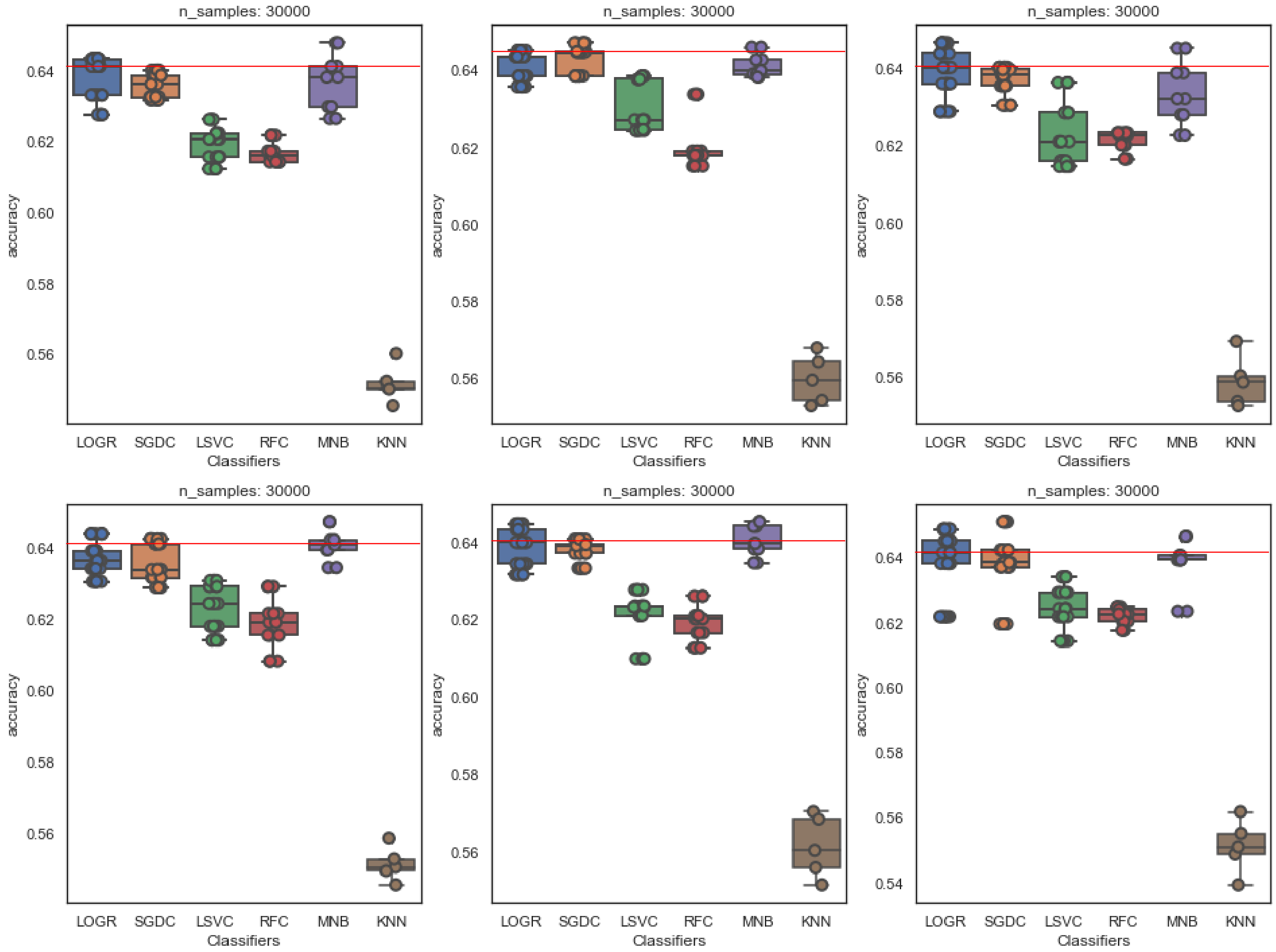
- **Evaluating Multiple Models**
- **Varying Sample Size**
- **Model 1: Parameter Selection, Model Fit**
- **Model 2: Parameter Selection, Model Fit**
- **Model 2: Fit Model**

- **Models Evaluated:**
  - Random Forest Classifier (RFC)
  - Multinomial Naïve Bayes (MNB)
  - Logistic Regression (LOGR)
  - Stochastic Gradient Descent Classifier (SGDC)
  - Linear Support Vector Classification (LSVC)
  - K Neighbors Classifier (KNN)

MODEL IMPLEMENTATION, RESULTS

Evaluating Multiple Models

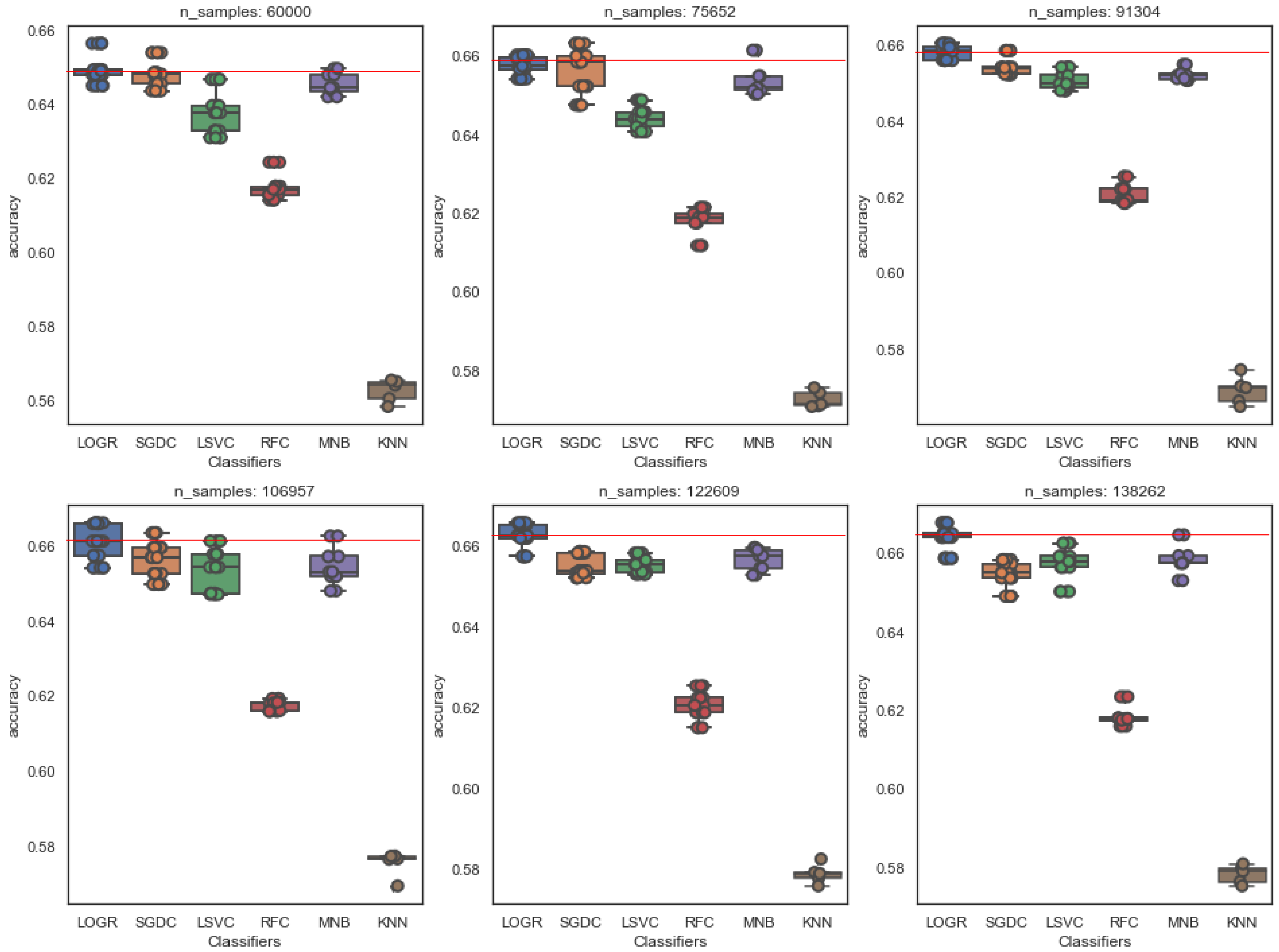
- **Modeling:**
  - 5-Fold Cross-Validation
  - 60/40 Train/Test Split
  - Randomized sample n = 30000
  - Default Parameters
  - Scored with *Accuracy*



MODEL IMPLEMENTATION, RESULTS

Varying Sample Size

- **Modeling:**
  - 5-Fold Cross-Validation
  - 60/40 Train/Test Split
  - Randomized sample n = 30000
  - Default Parameters
  - Scored with *Accuracy*



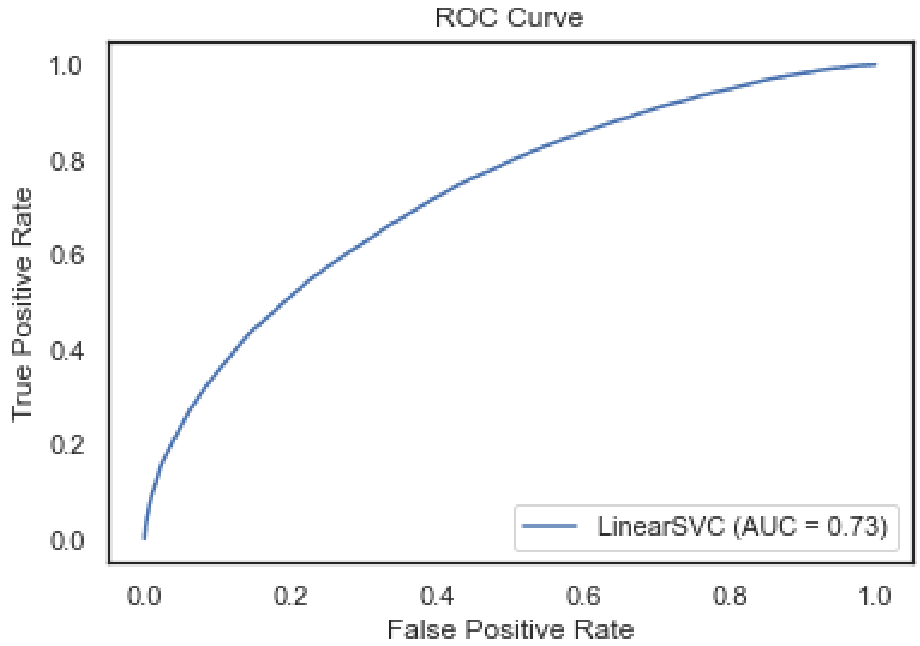
MODEL IMPLEMENTATION, RESULTS

- **Linear Support Vector Classification**
- **Parameter Selection:**
  - Grid Search
  - 10-Fold Cross-Validation
  - 70/30 Train/Test Split
  - Randomized sample n = 41479 (30% of Dataset)
  - Scored with *Accuracy*

- **Parameter Constants:**
  - *class_weight* = ‘balanced’
  - *max_iter* = ‘1000’

- **Parameter Grid:**
  - ***C*** across -5 to 4 (10 values across base 10 log space)

- **Best Fit:**
  - ***C*** = ‘0.1’




```
[[15637  7315]
 [ 6599 11928]]
              precision    recall  f1-score   support

         0.0       0.70      0.68      0.69     22952
         1.0       0.62      0.64      0.63     18527

    accuracy                           0.66     41479
   macro avg       0.66      0.66      0.66     41479
weighted avg       0.67      0.66      0.67     41479
```

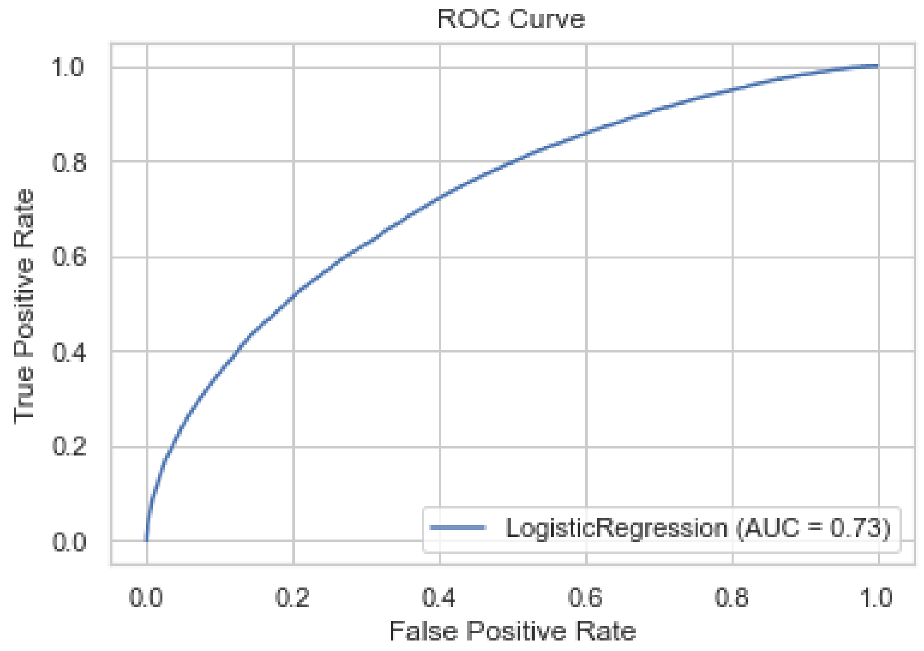
MODEL IMPLEMENTATION, RESULTS

- **Logistic Regression**
- **Parameter Selection:**
  - Grid Search
  - 10-Fold Cross-Validation
  - 70/30 Train/Test Split
  - Randomized sample n = 41479 (30% of Dataset)
  - Scored with *Accuracy*

- **Parameter Constants:**
  - *class_weight* = ‘balanced’
  - *max_iter* = ‘1000’

- **Parameter Grid:**
  - ***C*** across -5 to 4 (10 values across base 10 log space)
  - ***Solver*** across ‘liblinear’, ‘saga’

- **Best Fit:**
  - ***C*** = ‘0.1’
  - ***Solver***: ‘liblinear’




```
[[15594  7358]
 [ 6573 11954]]
              precision    recall  f1-score   support

         0.0       0.70      0.68      0.69     22952
         1.0       0.62      0.65      0.63     18527

    accuracy                           0.66     41479
   macro avg       0.66      0.66      0.66     41479
weighted avg       0.67      0.66      0.66     41479
```

CLOSING

- **Gender-Related Nouns**
  - boy list: *wife*
  - girl list: *husband*
- **Diction**
  - boy list: expletives, slang, job roles
  - girl list: different job roles
- **Spelling**
  - Conscious, unconscious spelling
  - boy list: higher frequency of both
  - girl list: less frequent
- Reflecting on **Context, Author Identity**

- **Future Improvements**
  - Addressing Processes Introducing Bias
  - Alternative Cleaning Processes
  - Improved Modeling Process
  - Improved Model Selection Process
  - Removing Redundancy in Model Selection

REFERENCES

Cohen, W. W. (2015). Enron Email Dataset. Retrieved from https://www.cs.cmu.edu/~./enron/.

Kantrowitz, M., Ross, B. (1994). *Names Corpus, Version 1.3*. Retrieved from http://www.cs.cmu.edu/afs/cs/project/ai-repository/ai/areas/nlp/corpora/names/

Prabhakaran, S. (2020). *Cosine Similarity – Understanding the math and how it works (with python codes).* Retrieved from https://www.machinelearningplus.com/nlp/cosine-similarity/.

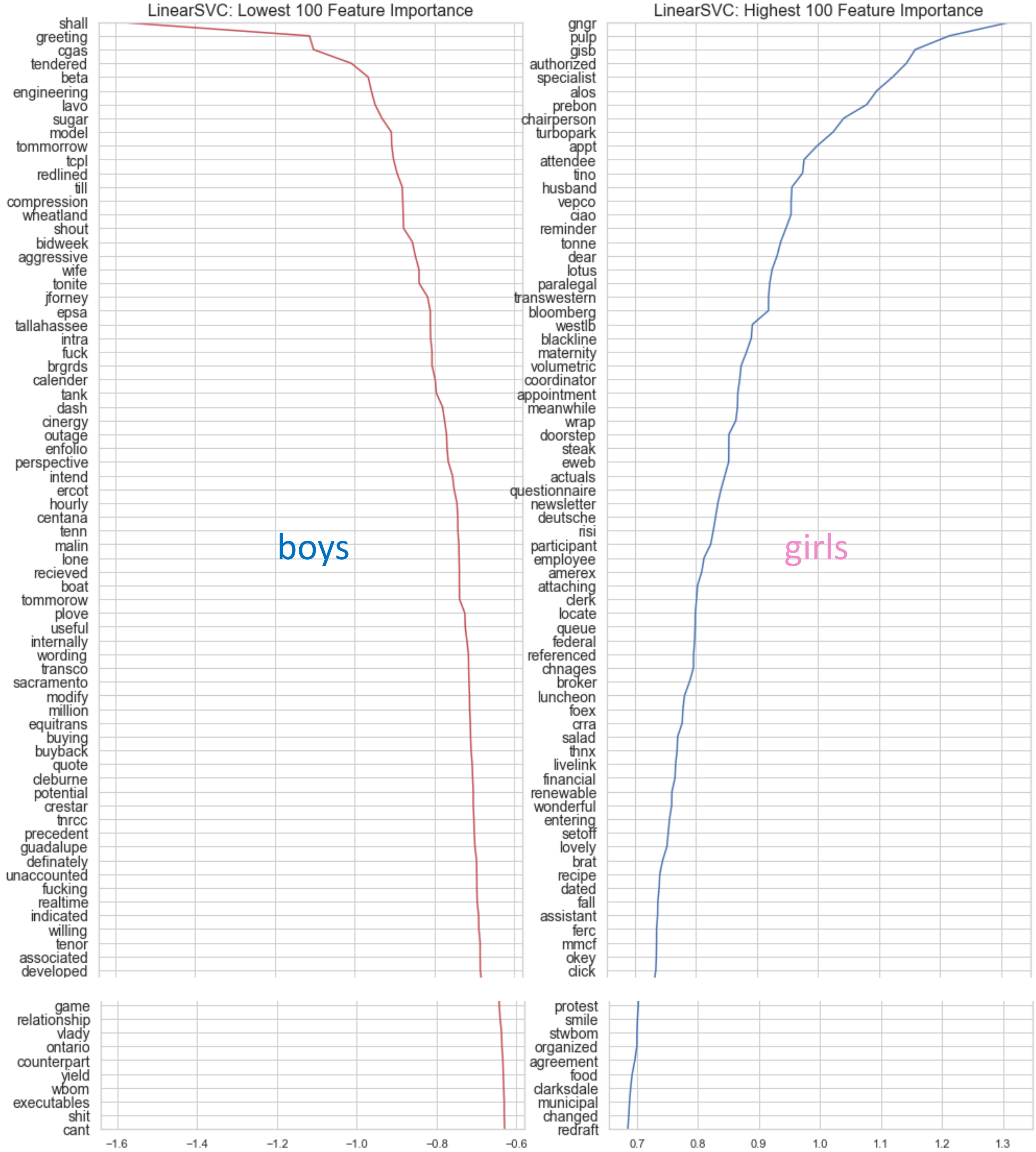
Scikit-learn. (2019a). *Sklearn.linear_model.SGDClassifier.* Retrieved from https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDClassifier.html

Scikit-learn. (2019b). *Sklearn.metrics.pairwise.cosine_similarity*. Retrieved from https://scikit-learn.org/stable/modules/generated/sklearn.metrics.pairwise.cosine_similarity.html.

Wikipedia. (2020). *Fundie*. Retrieved from https://en.wikipedia.org/wiki/Fundie

APPENDIX:

LINEAR SVC FEATURES



LOGISTIC REGRESSION FEATURES