Automated Fraud Detection

Using the Enron Email Corpus to Train Fraud Detection Models

By: Lin ZheQin

Table of Contents

Problem Statement:

Create machine models to predict corruption in emails using text mining

Stakeholders:

- Auditors/Regulators
- Board of Directors

Background

Infographic obtained from: "The Scandal The Failure of the Enron – Ravasqueira.com

Enron Email Dataset

- This data was originally made public by the Federal Energy Regulatory Commission during its investigation.
- Data has been downloaded from www.Kaggle.com
- Size: 1.3 GB
- The data contains more than 500,000 emails, retrieved from the user folders of 150 Enron employees
- These emails were sent by more than 20,000 unique email addresses.

	file	message
0	allen-p/_sent_mail/1.	Message-ID: <18782981.1075855378110.JavaMail.e...
1	allen-p/_sent_mail/10.	Message-ID: <15464986.1075855378456.JavaMail.e...
2	allen-p/_sent_mail/100.	Message-ID: <24216240.1075855687451.JavaMail.e...
3	allen-p/_sent_mail/1000.	Message-ID: <13505866.1075863688222.JavaMail.e...
4	allen-p/_sent_mail/1001.	Message-ID: <30922949.1075863688243.JavaMail.e...

Contents of a Sample Message

```
Message-ID: <13505866.1075863688222.JavaMail.evans@thyme>
Date: Mon, 23 Oct 2000 06:13:00 -0700 (PDT)
From: phillip.allen@enron.com
To: randall.gay@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: Randall L Gay
X-cc:
X-bcc:
X-Folder: \Phillip_Allen_Dec2000\Notes Folders\'sent mail
X-Origin: Allen-P
X-FileName: pallen.nsf

Randy,

 Can you send me a schedule of the salary and level of everyone in the
scheduling group.  Plus your thoughts on any changes that need to be made.
(Patti S for example)

Phillip
```

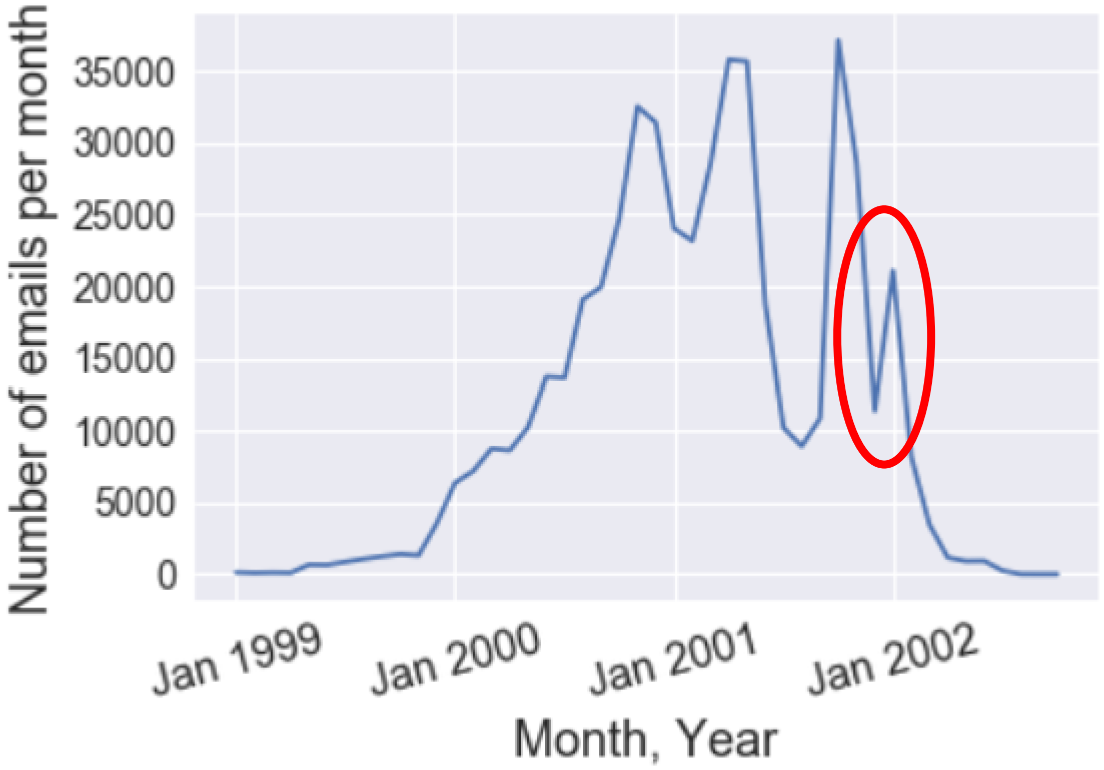
Final Data Frame

Message-ID	user	Date	From	To	Subject	X-From	X-To	X-cc	X-bcc	X-Folder	X-Origin	X-FileName	content
<18782981.1075855378110.JavaMail.evans@thyme>	allen-p	2001-05-14 23:39:00	phillip.allen@enron.com	(tim.belden@enron.com)		Phillip K Allen	Tim Belden <Tim Belden/Enron@EnronXGate>			\Phillip_Allen_Jan2002_1\Allen, Phillip K.\'Se...	Allen-P	pallen (Non-Privileged).pst	Here is our forecast\n\n
<15464986.1075855378456.JavaMail.evans@thyme>	allen-p	2001-05-04 20:51:00	phillip.allen@enron.com	(john.lavorato@enron.com)	Re:	Phillip K Allen	John J Lavorato <John J Lavorato/ENRON@enronXg...			\Phillip_Allen_Jan2002_1\Allen, Phillip K.\'Se...	Allen-P	pallen (Non-Privileged).pst	Traveling to have a business meeting takes the...
<24216240.1075855687451.JavaMail.evans@thyme>	allen-p	2000-10-18 10:00:00	phillip.allen@enron.com	(leah.arsdall@enron.com)	Re: test	Phillip K Allen	Leah Van Arsdall			\Phillip_Allen_Dec2000\Notes Folders\'sent mail	Allen-P	pallen.nsf	test successful. way to go!!!
<13505866.1075863688222.JavaMail.evans@thyme>	allen-p	2000-10-23 13:13:00	phillip.allen@enron.com	(randall.gay@enron.com)		Phillip K Allen	Randall L Gay			\Phillip_Allen_Dec2000\Notes Folders\'sent mail	Allen-P	pallen.nsf	Randy,\n\n Can you send me a schedule of the s...
<30922949.1075863688243.JavaMail.evans@thyme>	allen-p	2000-08-31 12:07:00	phillip.allen@enron.com	(greg.piper@enron.com)	Re: Hello	Phillip K Allen	Greg Piper			\Phillip_Allen_Dec2000\Notes Folders\'sent mail	Allen-P	pallen.nsf	Let's shoot for Tuesday at 11:45.

Exploratory Data Analysis (EDA)

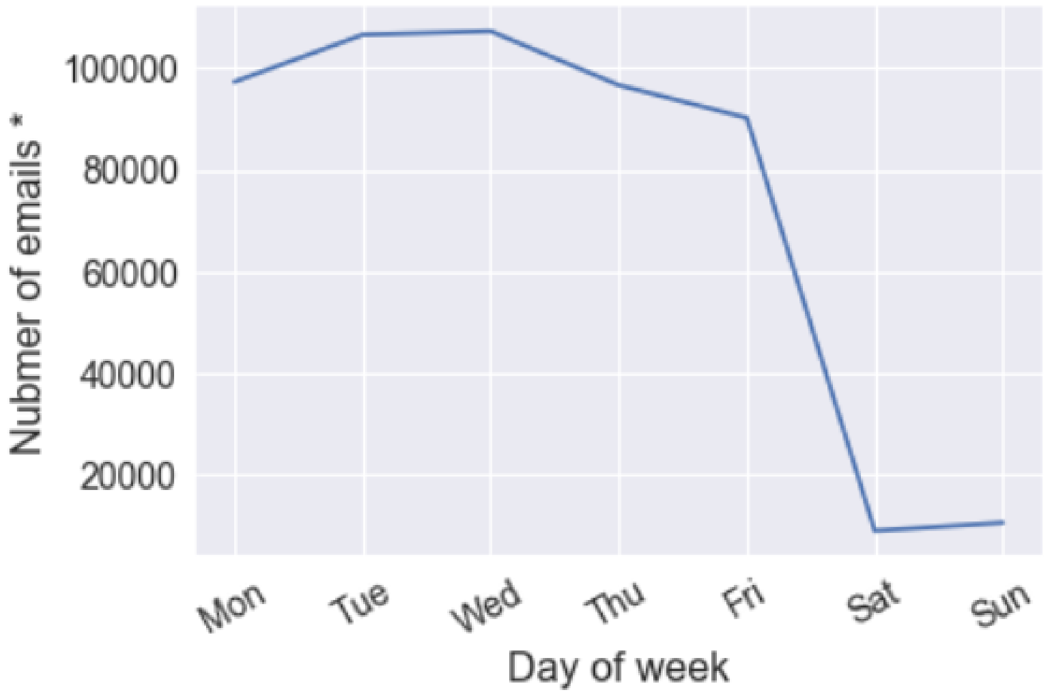
- The use of emails at Enron picked up in 1999 and increased steadily during 2000s.

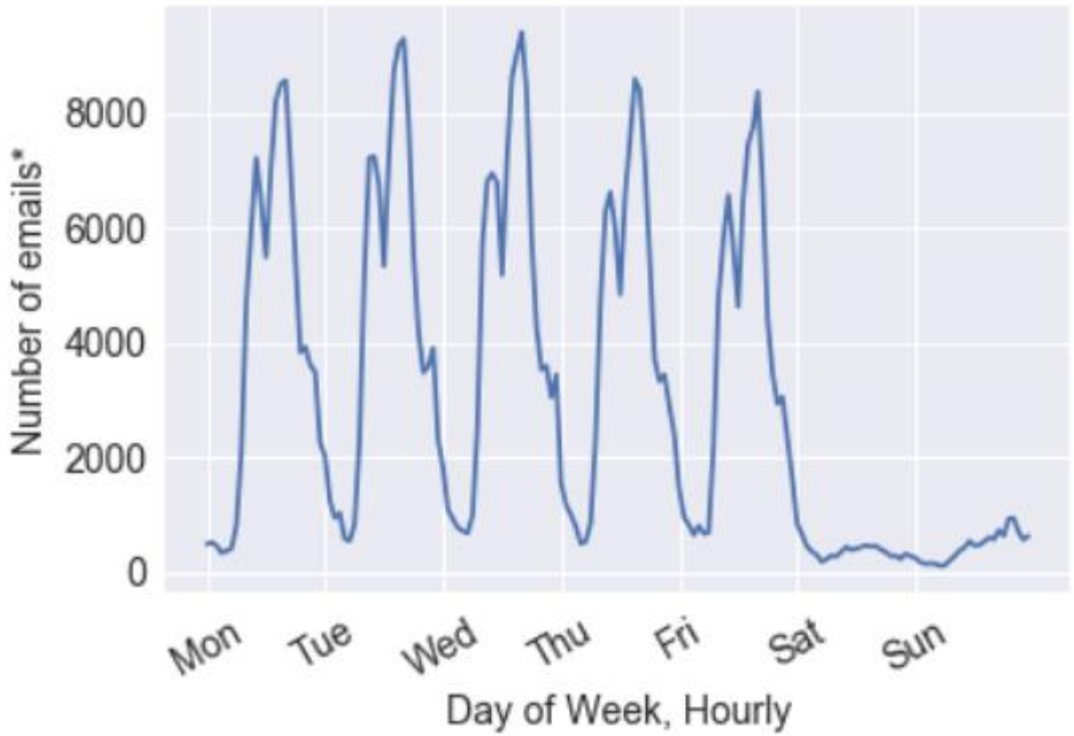
- In 2001, the year that Enron collapsed, there was a sudden drop in the volume of emails during summer months followed by a sharp peak in fall and a steady drop after the bankruptcy.



Exploratory Data Analysis (EDA)

Hourly and daily email volume





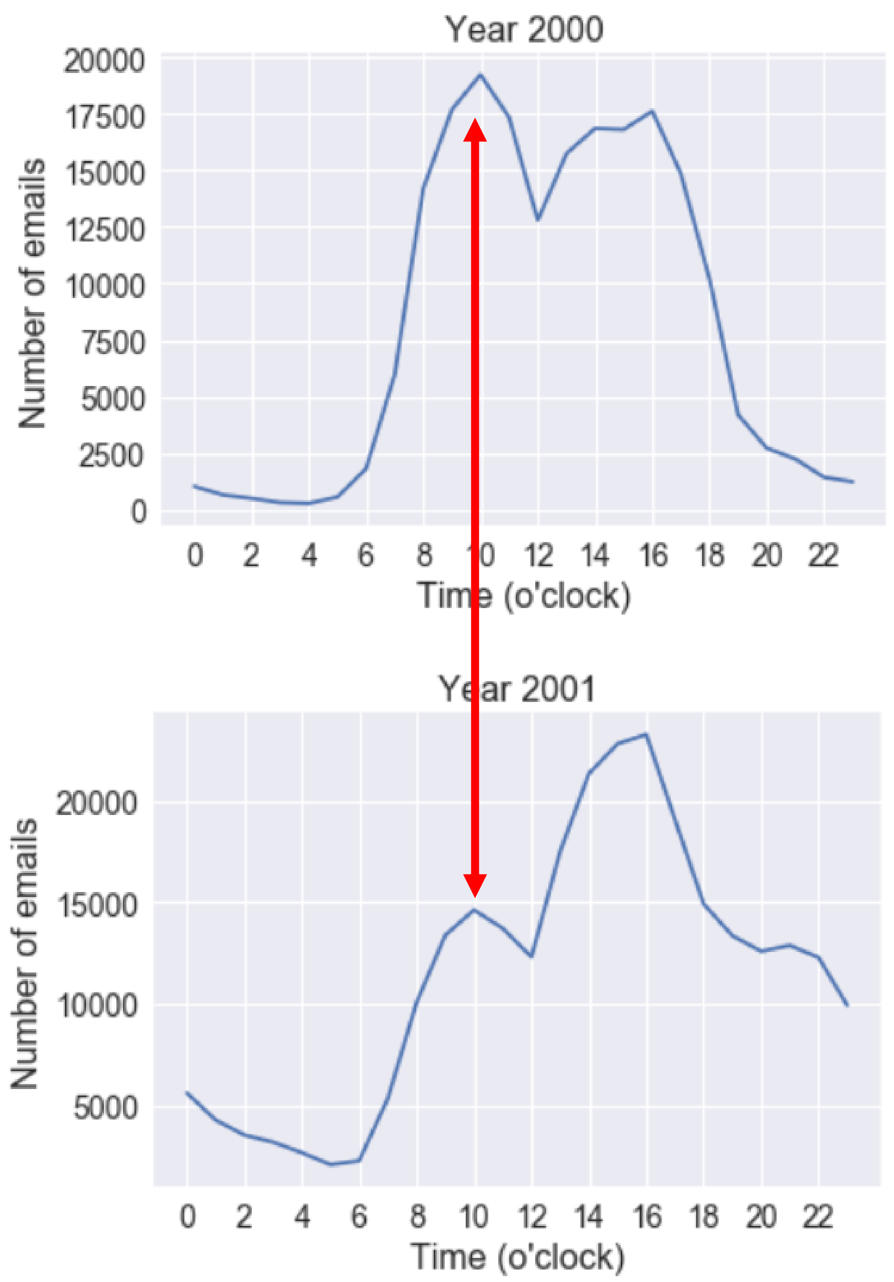
* The y axis represent the total email count during the livelihood of Enron, not the average daily volume

Hypothesis Testing

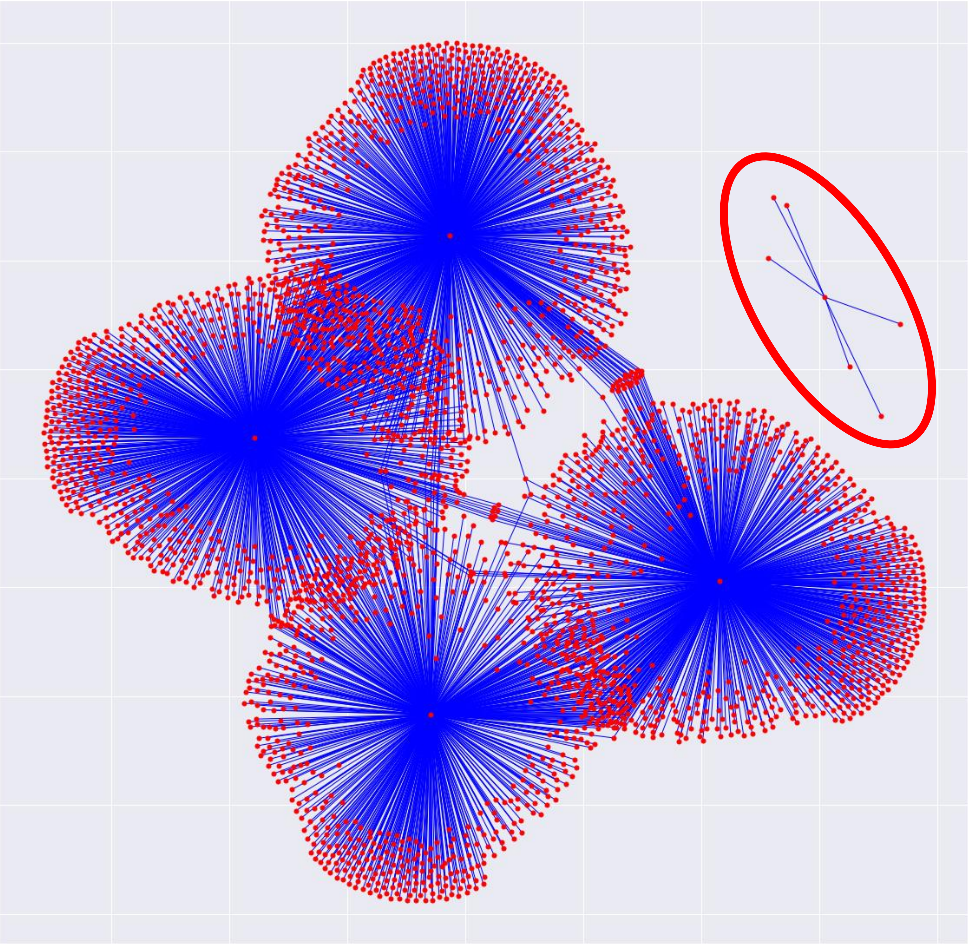
- **Question:** Is there any significant difference in the email volume between 8 to 10 am and 3 to 5 pm in 2001, excluding weekends?

- **Motivation:** By looking at the hourly email volume in 2000 and 2001 we can notice a difference between the volume of emails in the mornings and in the afternoons in 2001. 2001 is the year that Enron collapsed.

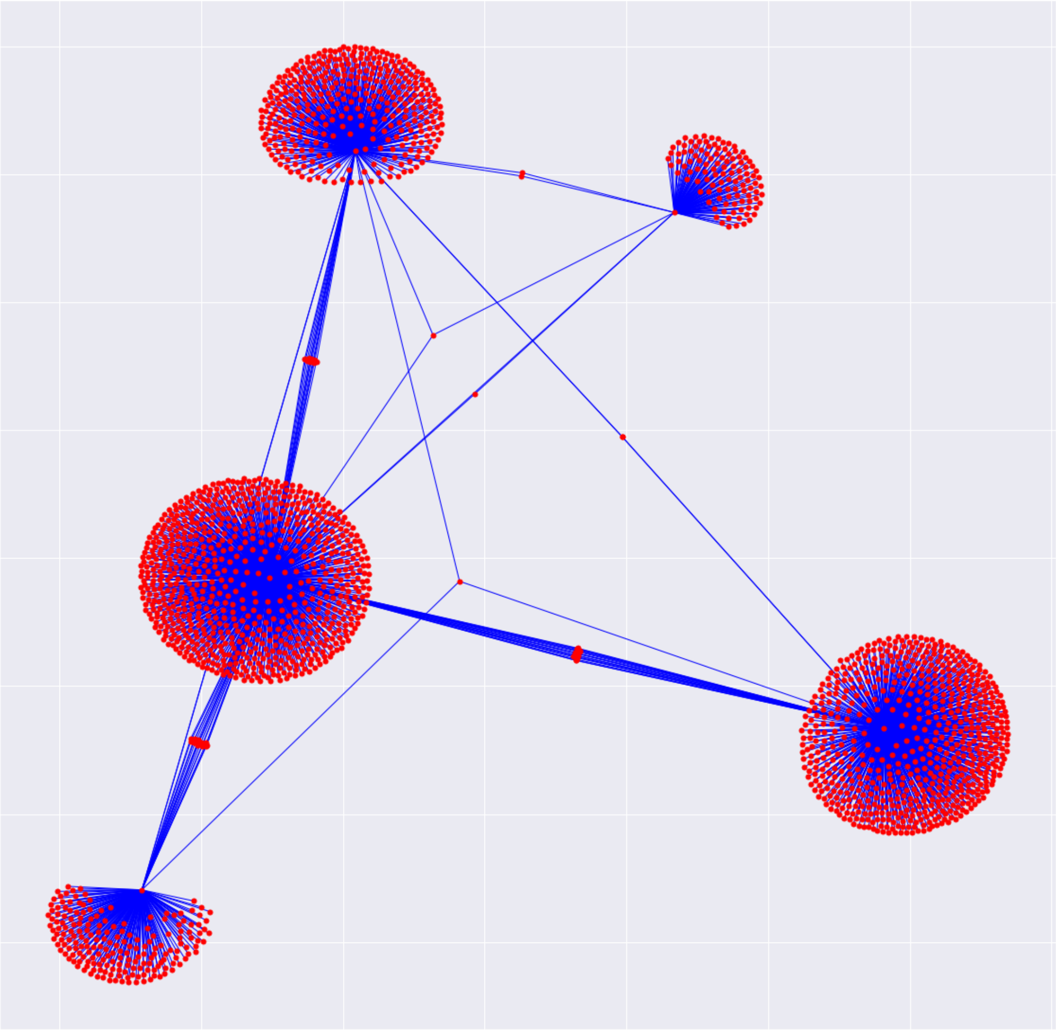
  One may interpret a higher volume of emails in the afternoons vs. mornings as a ***sign of procrastination*** or ***lack of interest by employees.***



Network Visualization



Network of top 5 email senders



Network of the next 5 email senders

Interesting observation - Outlier

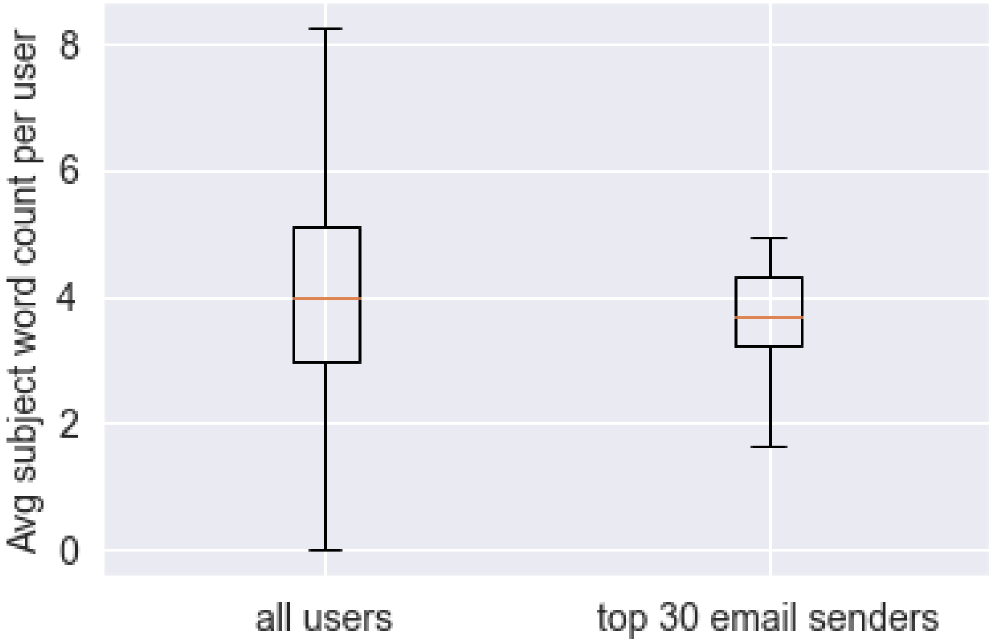
- Outlier - 80% of emails sent to oneself

- Review of email content – as a lab notebook to keep a log of work.

```
From
frozenset({'jeff.dasovich@enron.com'})      794
frozenset({'kay.mann@enron.com'})           647
frozenset({'pete.davis@enron.com'})           7
frozenset({'sara.shackleton@enron.com'})    826
frozenset({'vince.kaminski@enron.com'})     794
Name: Recipient_1, dtype: int64
```

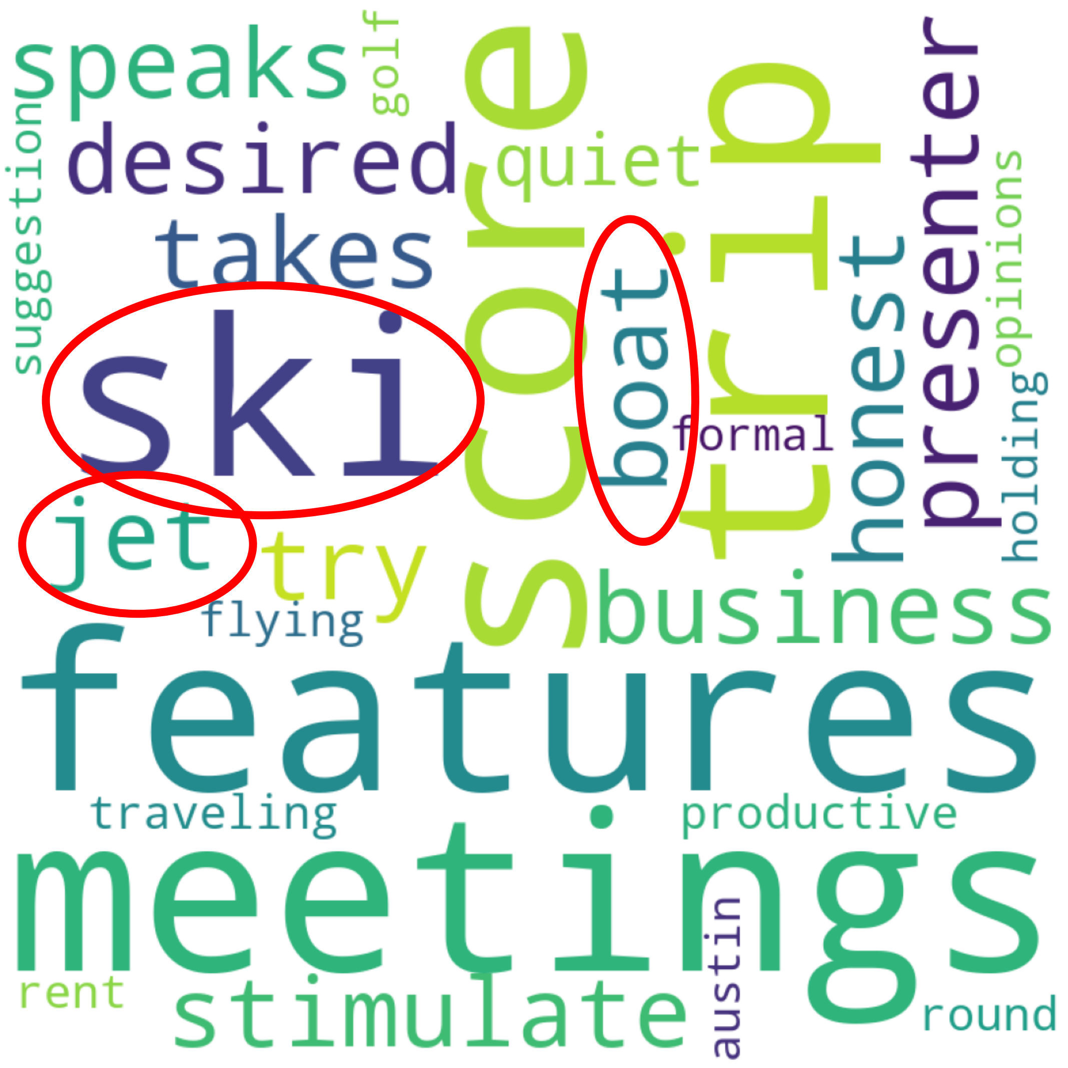
Inferential Statistics - Question 1:

- Did people who send a lot of emails write shorter emails?
- Comparing top 30 email senders with all email senders in terms of subject and contents word count.
- From the boxplot it is clear that there is no significant difference in the subject and contents word count in these two groups.



Word Cloud

- The figure was produced using the Word Cloud library
- Top words – emails topics more around 'play' than work?



Modeling & Evaluation

- Unsupervised Learning – K-Means Clustering
- 2 clusters
  - Batch size 500
  - 100 iterations
- Top Features
  - Non-work related

	features	score
0	meetings	0.396321
1	trip	0.303915
2	ski	0.279288
3	business	0.260357
4	takes	0.207651
5	presenter	0.174026
6	try	0.165728
7	stimulate	0.165694
8	speaks	0.150587
9	jet	0.146178
10	boat	0.144293
11	desired	0.139921
12	honest	0.139807
13	quiet	0.137759
14	productive	0.136461
15	rent	0.133112
16	flying	0.130266
17	traveling	0.124373
18	golf	0.123238
19	suggestion	0.121763
20	formal	0.116549
21	opinions	0.115387
22	round	0.110326
23	holding	0.108622
24	austin	0.107847

Modeling & Evaluation

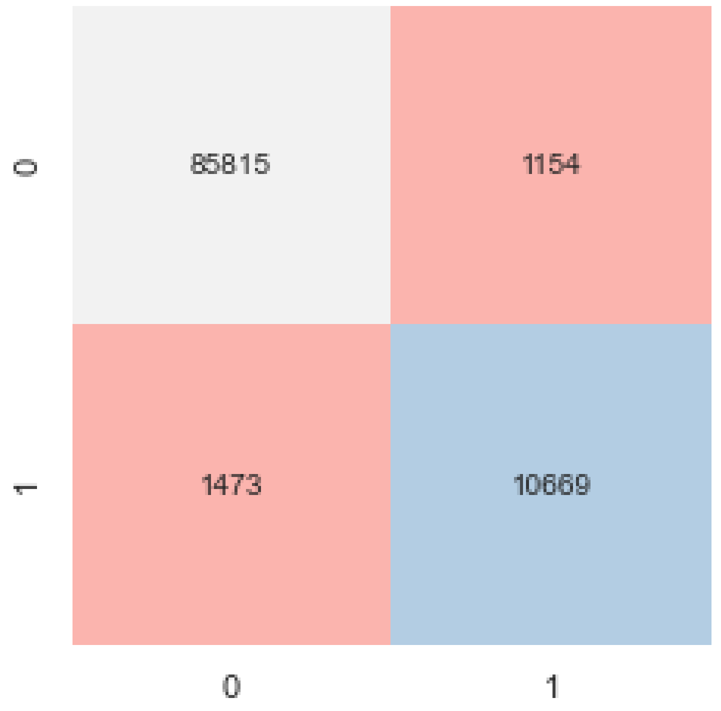
- **Supervised Learning** – *K-Nearest Neighbours*
- Test size – 20%
- Words in emails separated into different clusters
    - Non-work related words vs common words
- Computation is deferred until classification – desired for large dataset

```
[[85815  1154]
 [ 1473 10669]]
              precision    recall  f1-score   support

           0       0.98      0.99      0.98     86969
           1       0.90      0.88      0.89     12142

    accuracy                           0.97     99111
   macro avg       0.94      0.93      0.94     99111
weighted avg       0.97      0.97      0.97     99111

0.9734943649039965
```



Conclusion / Next Steps

- Application of trained model in Fraud and Risk Management
  - Utilize Machine Learning as first indicator of red flags
    - ✓ Highlight people at risk of committing fraud – more efficient for auditors to do in-depth review
- Recommendations
  - Getting access to computing resources to run the algorithm on full dataset
  - Deeper review of language used in emails

Questions?