Monitoring Financial Advice Files Regtech Initiative

Presentation 5

Flexprod Industries

Glossary

- SoA – Statement of Advice
- ML – Machine Learning
- POS tag – Part of Speech tagging
- NLP – Natural Language processing
- TF-IDF – Term Frequency, Inverse-Document Frequency

My background

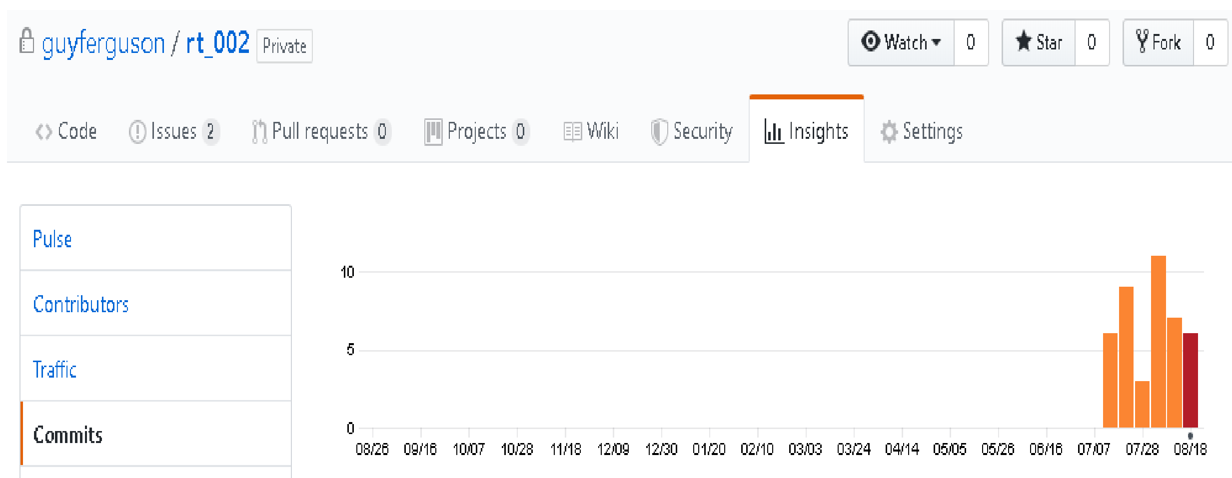
- Work in the financial sector – now super, previously banking
- Find answers for a variety of data problems
- Recent IT graduate, software development with focus on AI and ML

Showcase innovative technology:

- With your own product
- Build your own application
- Deliver presentations / ideas / proofs of concept

Excuses slide

First commit – July 17



<guyferguson@tp	2019-08-19 23:29:40
<guyferguson@tp	2019-08-19 00:02:23
<guyferguson@tp	2019-08-18 20:56:29
<guyferguson@tp	2019-08-18 10:46:15
<guyferguson@tp	2019-08-18 10:45:02
<guyferguson@tp	2019-08-18 10:37:55
<guyferguson@tp	2019-08-18 00:06:56
<guyferguson@tp	2019-08-16 12:41:32
<guyferguson@tp	2019-08-14 22:05:04
<guyferguson@tp	2019-08-14 19:53:43
<guyferguson@tp	2019-08-13 22:34:06
<guyferguson@tp	2019-08-12 21:59:51
<guyferguson@tp	2019-08-11 23:44:14
<guyferguson@tp	2019-08-11 00:15:24
<guyferguson@tp	2019-08-09 23:54:21
<guyferguson@tp	2019-08-09 23:33:09
<guyferguson@tp	2019-08-08 23:19:36
<guy.ferguson@l	2019-08-08 17:20:32
<guyferguson@tp	2019-08-07 23:43:54
<guy.ferguson@l	2019-08-07 17:37:49

…all commits between 6PM and 1AM, (when the best code is written)

Solution approach

- Python 3.7 (5 week dev not achievable with other languages)
  - Libraries:
    - Nltk (natural language tool kit) for text-pre-processing
    - Tika for pdf/Word parsing
    - Scikit, pandas, numPy for the ML
- Data
  - ASIC dataset (20 clients)
  - 3 real-world SoA instances + RG90 Appendix 2 SoA

Supervised learning + Classification problem

File details		
Number	**Client Name**	**Goal disclosure**
1	Wilma Flintstone	PASS
2	Keanu Roves	PASS
3	Roger & Diana Rabbet	PASS
4	Sean Conneray	PASS
5	Timothy Dixon	PASS
6	Grace Codd	FAIL
7	Pierce Brown & Paula Brown	PASS
8	Mary Poppins	PASS
9	John and Jane Wick in their capacity as trustees of the Parabellum SMSF	PASS
10	Cindy Rella	PASS
11	Cruella De Ville	PASS
12	Dr. Stephen Strange	PASS
13	Mrs Ygritte Snow & Mr John Snow	PASS
14	Daniel Cray & Eva Cray	PASS
15	Mon Gustave	PASS
16	Bruce Li	FAIL
17	Anthea Saint & Lou Burns	PASS
18	Jim Jones	PASS
19	LeBron Jones	PASS
20	Katniss Ye	PASS
21	Bruce Bogtrotter	PASS
22	Jack Hill	PASS
23	Brad Black (ASIC sample doc)	PASS
24	Michael Williams	PASS

First mis-steps

- Initially I amalgamated all documents per client into one set of words.
- The usual ML text pre-processing steps:
  - Remove punctuation
  - Remove ‘stop words’ (‘the’,’and’,’a’,’as’…)
  - Create stemmas and lemmas, pos_tag
  - Create n-grams:

```
('date', 'complet'), ('complet', 'financi'), ('financi', 'servic'), ('servic', 'guid'),
```

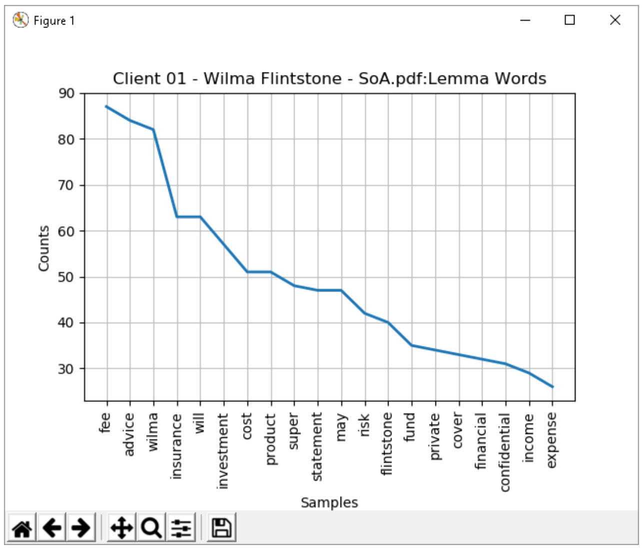
Data extraction

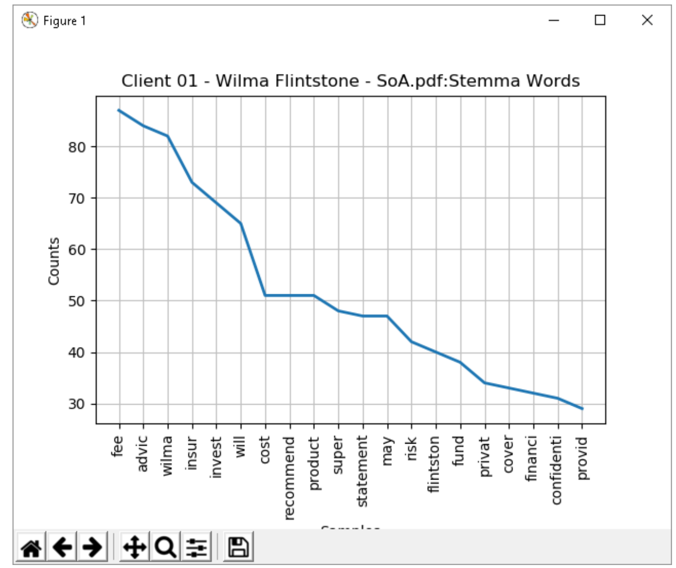
- Finishes with a client –data extracted and processed – 4 seconds all up for all documents

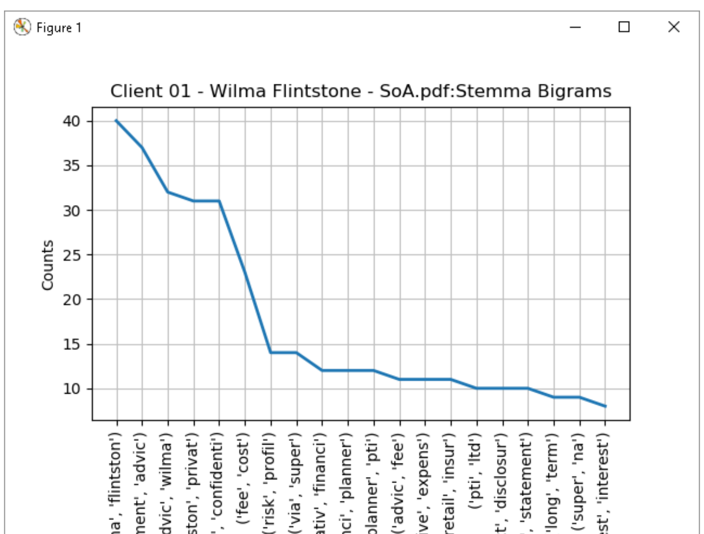
F:\Python37\python.exe

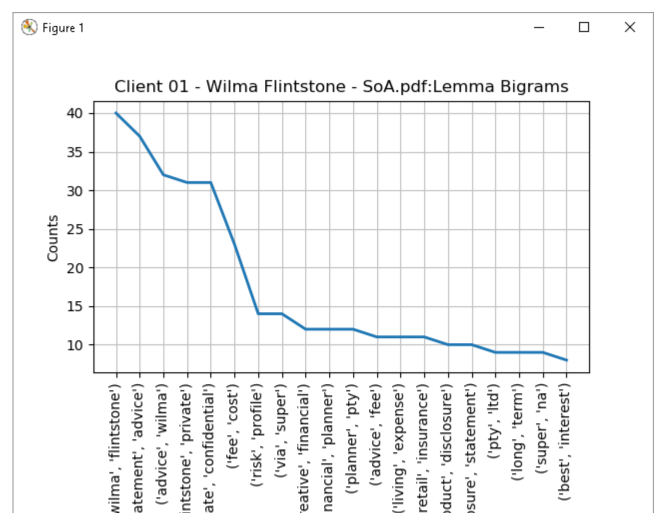
```
102.91 .....finished with : Client 02 - Keanu Roves - File Note.pdf
102.91 ....looking for doc 6
----------------------
       ....evaluating file H:\rt_files\Data\Client 02 - Keanu Roves\Client 02 -
Keanu Roves - SoA.pdf
       ....read doc
       ....read ABN  33 333 333 333
       ....read AFSL  123456
       ....Stripped punctuation and numerics    :69946
       ....POS tagged
       ....Tokenised                            :10444
       ....Stops removed                        :4228
       ....Lemma set created                    :6216
       ....Lemma Bigram set created             :6215
       ....Stemma set created                   :6216
       ....Stemma Bigram set created            :6215
105.20 ....finished creating FinAdviceDoc object
     ....Added tokenised words to client 7590
     ....Added bigrams to client 7587
105.69 .....FinAdviceDoc words:           76317
105.69 .....Table count:         0
105.69 .....finished with : Client 02 - Keanu Roves - SoA.pdf
105.69 ....looking for doc 7
----------------------
----------------------
```

Getting data familiar – Flintstone SoA lemma and stemma unigrams and bigrams

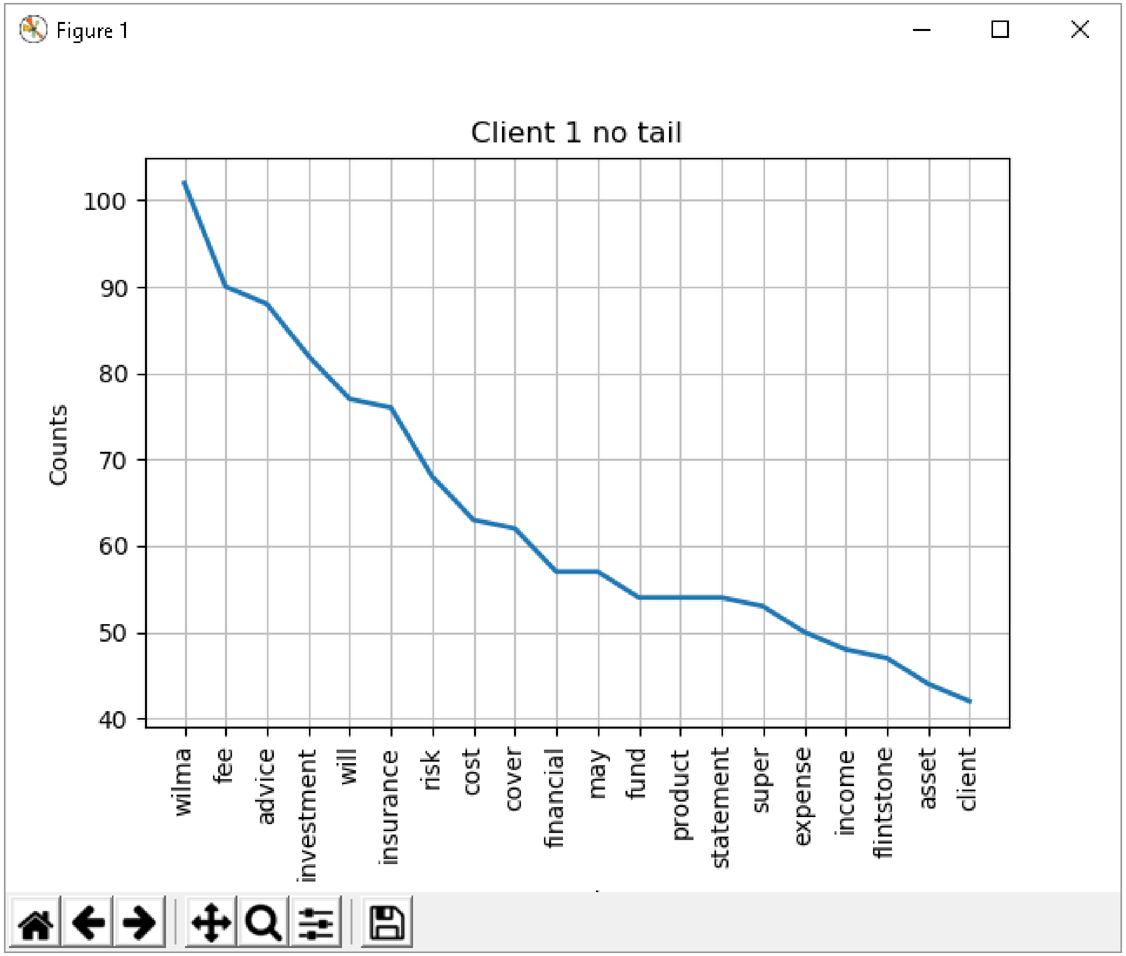








All client files combined - Flintstone



Most frequent words:

wilma : 104
fee : 90
advice : 88
investment : 82
will : 77

Mis-steps realised

Storing compliance risks :

```
# doc_type -   IQ   SOA FN FFR
self.risks = [{'id':1,'risk':'goals and objectives not included', 'weight':100,'doc_type': 'SOA'},
              {'id':2,'risk':'legislative warnings not included', 'weight':100,'doc_type': 'SOA'},
              {'id':3,'risk':'ABN not quoted', 'field':'self.ABN','operator':'!=\'Unknown\'','weight':100
              {'id':4,'risk':'AFSL not quoted','field':'self.AFSL', 'operator':'!=\'Unknown\'','weight':1
```

Potential compliance risks

- For product replacement advice, the risk that the Statement of Advice (SoA) does not include the requisite information required by legislation (having regard to relevant circumstances)
- The risk that the client's goals & objectives are not clearly stated in the SoA

Compliance Risk tests: 3 and 4: Run 22:27:52.930872

Assessing risks ABN not quoted,AFSL not quoted

Client ID:1: Name: Wilma Flintstone ABN: 22 222 222 222 AFSL: 654321______________ [Risk 3: PASS] RG 90.27(c),s947B(2)(c) [Risk 4: PASS] R

Client ID:2: Name: Keanu Roves ABN: 33 333 333 333 AFSL: 123456____________________ [Risk 3: PASS] RG 90.27(c),s947B(2)(c) [Risk 4: PASS]

Client ID:3: Name: Roger & Diana Rabbet ABN: 33 333 333 333 AFSL: 123456_________ [Risk 3: PASS] RG 90.27(c),s947B(2)(c) [Risk 4: PASS]

Client ID:4: Name: Sean Conneray ABN: 33 333 333 333 AFSL: 123456_________________ [Risk 3: PASS] RG 90.27(c),s947B(2)(c) [Risk 4: PASS]

Client ID:5: Name: Timothy Dixon ABN: 33 333 333 333 AFSL: 123456_________________ [Risk 3: PASS] RG 90.27(c),s947B(2)(c) [Risk 4: PASS]

Client ID:6: Name: Grace Codd ABN: 44 444 444 444 AFSL: 555555_____________________ [Risk 3: PASS] RG 90.27(c),s947B(2)(c) [Risk 4: PASS

Client ID:7: Name: Pierce & Paula Brown ABN: 33 333 333 333 AFSL: 123456_________ [Risk 3: PASS] RG 90.27(c),s947B(2)(c) [Risk 4: PASS]

Concordances of ‘goal/objective’ synonyms.

Flintstone – a PASSing goal disclosure

Client ID:1: Name: Wilma Flintstone ABN: 22 222 222 222 AFSL: 654321____________ [Risk 3: PASS] RG 90.27(c),s947B(2)(c) [Ri

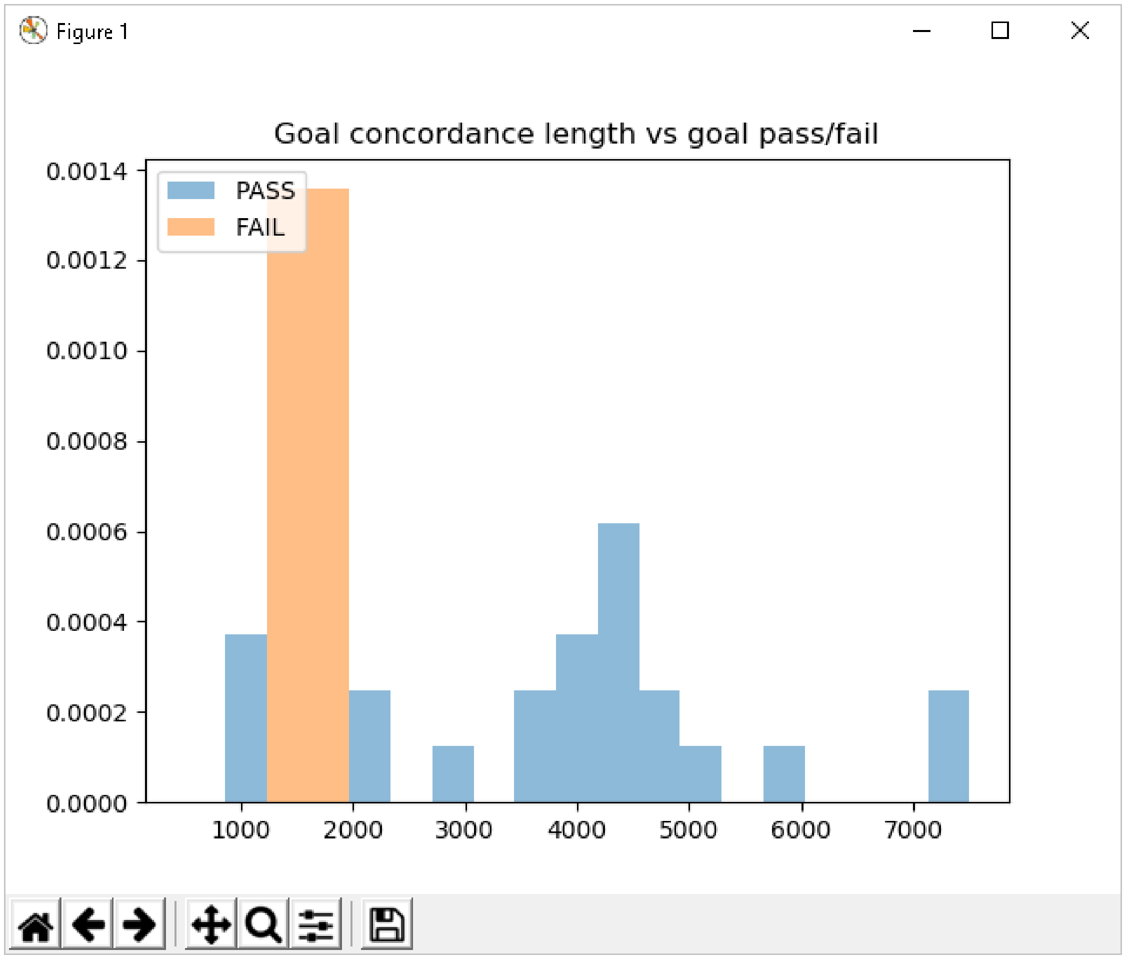
- **plan financial strategy meet lifestyle goal statement advice wilma flintstone private confidential content important thing need kno cost disclosure action required next step authority proceed acknowledgement declaration appendix current**
- **thing need know situation circumstance goal collected information can provide appropriate advice circumstance meet goal advic information incomplete incorrect may affect appropriateness advice therefore important use let u know case can review**
- **provide appropriate advice circumstance meet goal advice based information provided information incomplete incorrect may affe important use let u know case can review revise recommendation thing consider important understand risk associated advice**
- **confidential scope advice told u goal wilma goal foundation advice important thing u consider giving advice told u planning retire track planning work parttime retire like u review insurance prefer**
- **advice told u goal wilma goal foundation advice important thing u consider giving advice told u planning retire year want review parttime retire like u review insurance prefer diversified actively**
- **meeting earlier ensure track meeting goal create wealth future statement advice wilma flintstone private confidential risk profile strategy investment will help achieve goal developed analysing risk capacity tolerance**
- **strategy investment will help achieve goal developed analysing risk capacity tolerance appetite risk capacity refers extent can en fall value asset loss capital influenced factor reliance income investment rely another source**
- **relying investment income derived support goal ie starting pension fund retirement le time available investment recover generally risk le time recover adverse event investment generally categorised following short term year**
- **volatility willing accept investing achieve goal example can seen performance fund returned risk appetite amount type risk willin appetite may vary change time example may prefer invest le volatile**
- **type risk willing take achieve goal depending goal risk appetite may vary change time example may prefer invest le volatile inves provide living expense using combination following can create investment**
- **willing take achieve goal depending goal risk appetite may vary change time example may prefer invest le volatile investment like living expense using combination following can create investment portfolio appropriate**

Codd – a FAILing goal disclosure

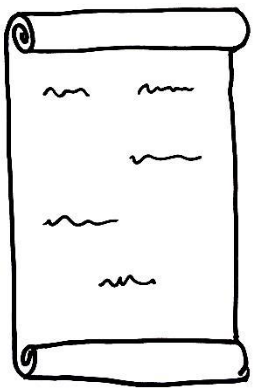
Client ID:6: Name: Grace Codd ABN: 44 444 444 444 AFSL: 555555________________ [Risk 3: PASS] I

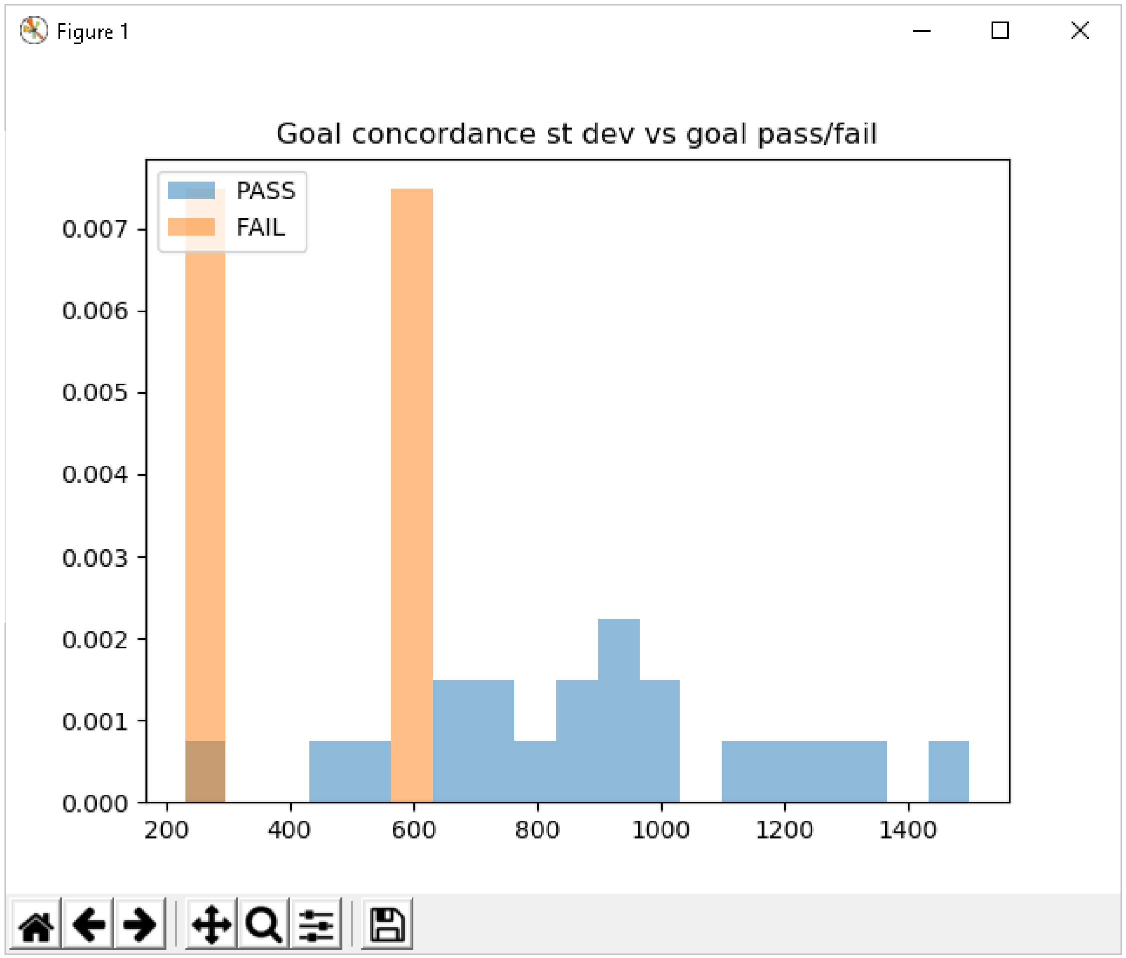
- advice respect overall financial circumstance goal happy review area within scope advice next review w superannuation advice recommendation happy way fashion superannuation fund performing wish recei
- grace summarised recommendation help achieve goal wealth creation invest geared equity recommend growth ozzie share option within krypton capital protected portfolio product estate planning recommen
- expectation feel investment risk considered goal wanted achieve timeframes involved recommend appro type investment risk profile type investor investment timeframe likelihood negative return indicative ret
- level investment risk order meet goal generate high return longterm addition expecting retire turn year timeframe goal objective grace please note investing risk profile likely result higher
- match risk tolerance investment timeframe goal objective grace please note investing risk profile likely higher allocation defensive asset confirmed u understand accept level risk high growth risk
- solution feel advice appropriate achieve goal recommendation grace recommended invest selection aust within krypton capital protected portfolio krypton facility allows borrow required capital via investment
- soon possible ensure wealth creation objective will meet insurance grace told u wish insurance reviewe includes strategy borrow invest highly recommend insurance reviewed soon possible potentially underin
- risk tolerance investment timeframe goal objective grace please note investing risk profile likely result l allocation defensive asset confirmed u understand accept level risk high growth risk profile
- financial position accumulate wealth line objective will able claim tax deduction interest expense will be interest expense exceeds assessable income received krypton investment may able use excess interest ex
- current personal financial position need objective understand information incomplete inaccurate advice statement product recommended within soa additional information listed soa applicable projection assu indicative

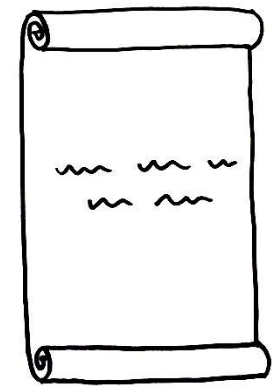
Feature extraction



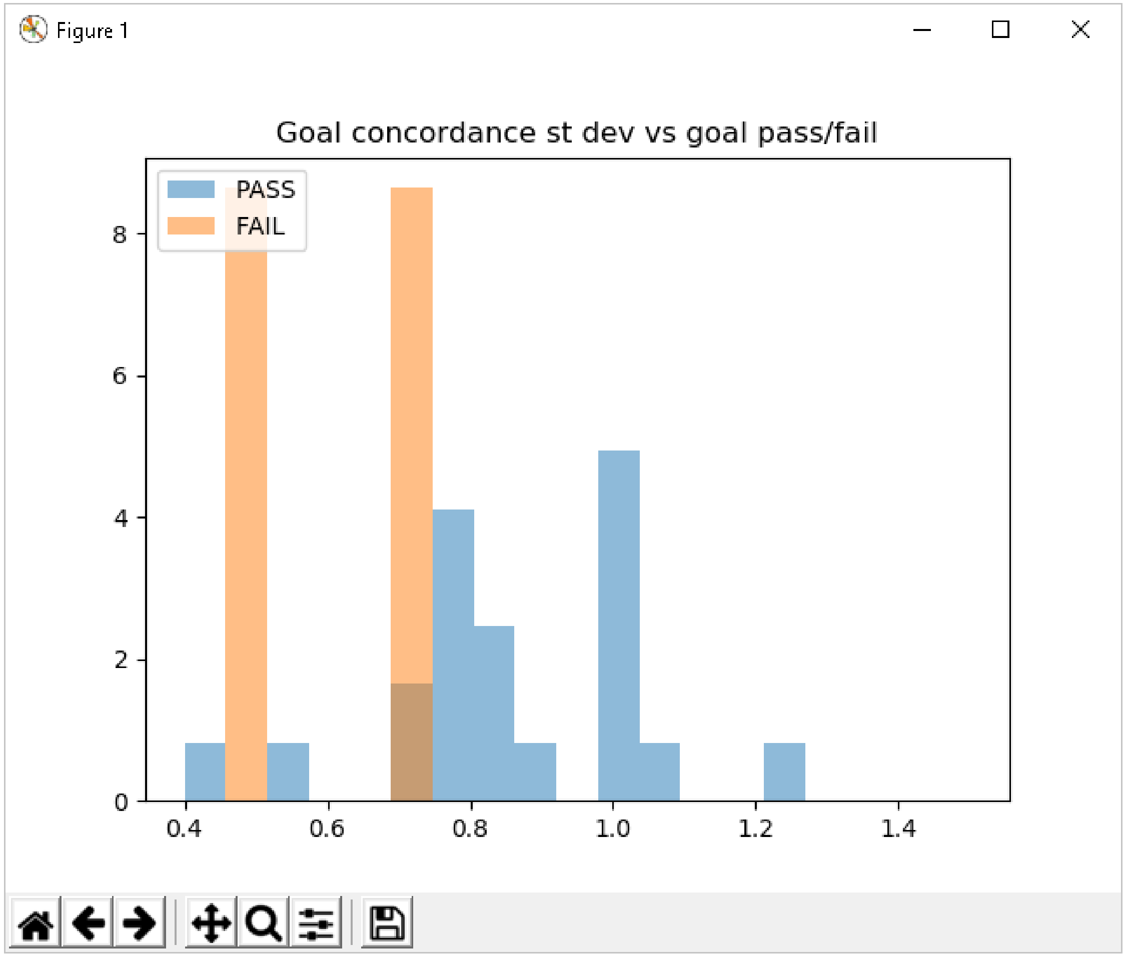
Standard deviation – measure of difference from mean



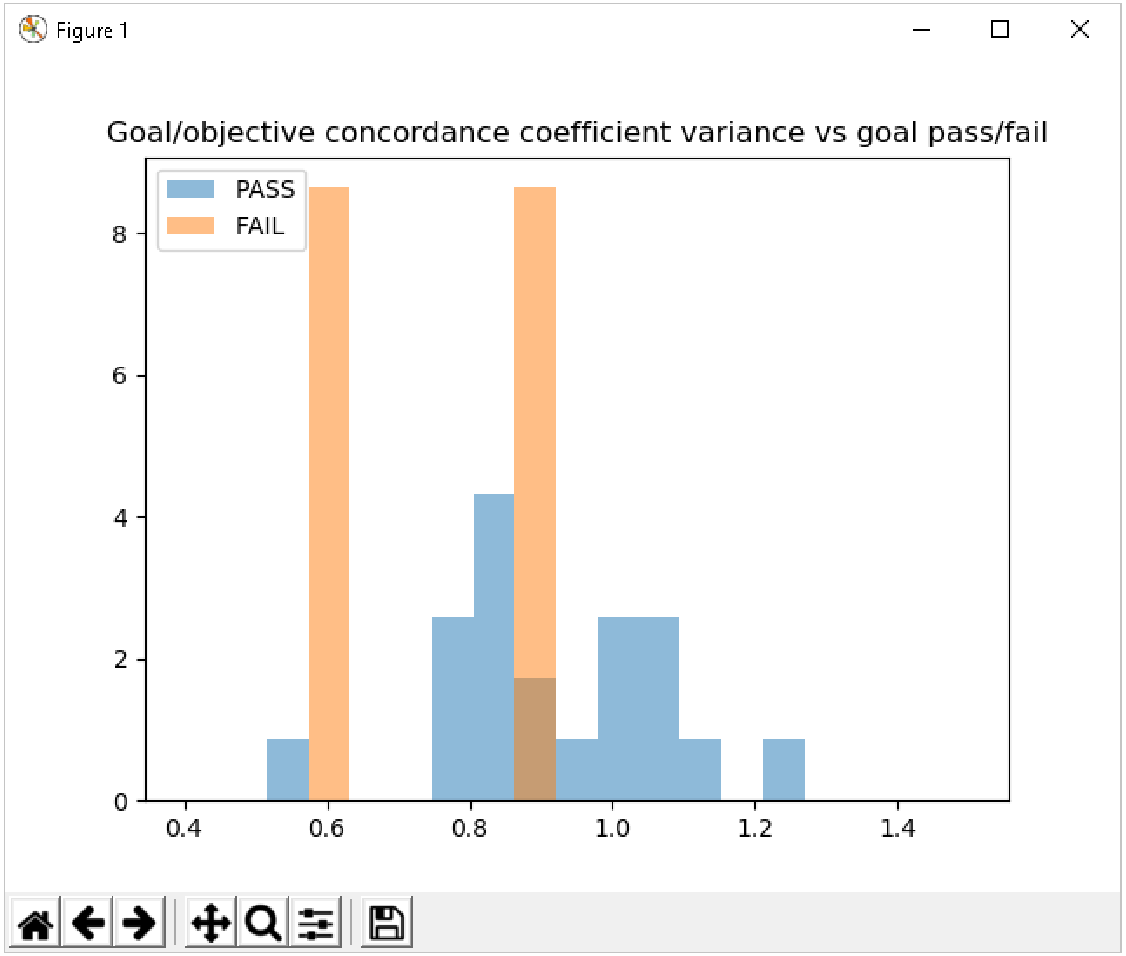




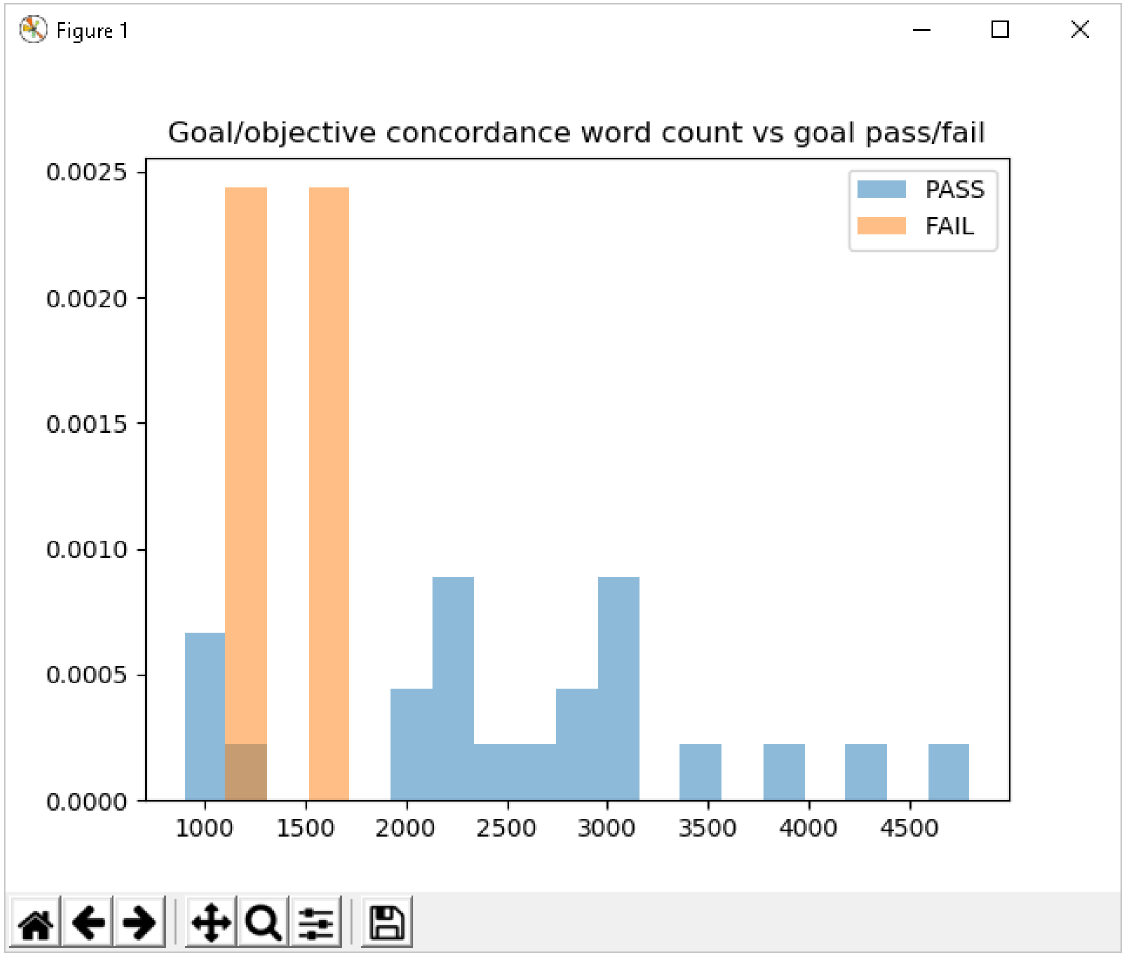
Normalised Standard deviation



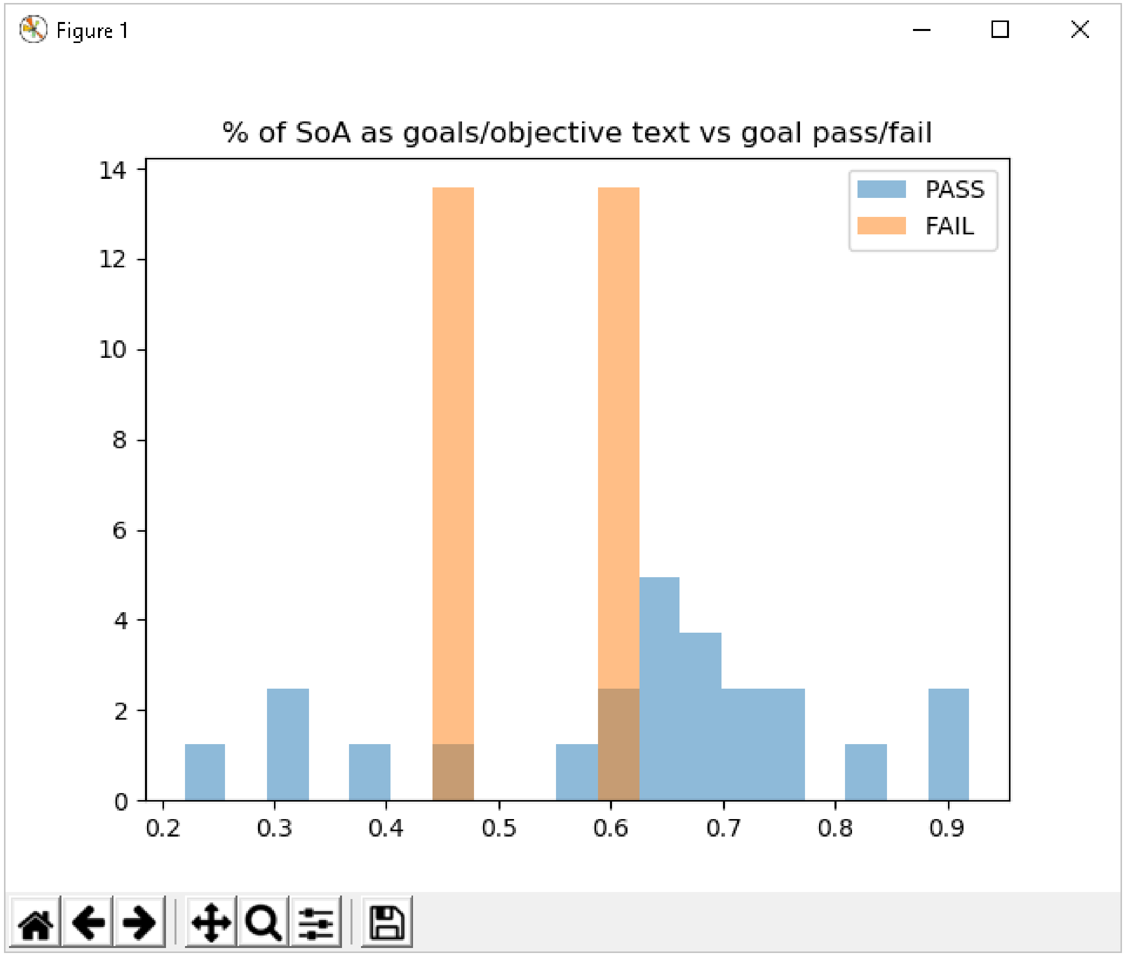
Coefficient of variation



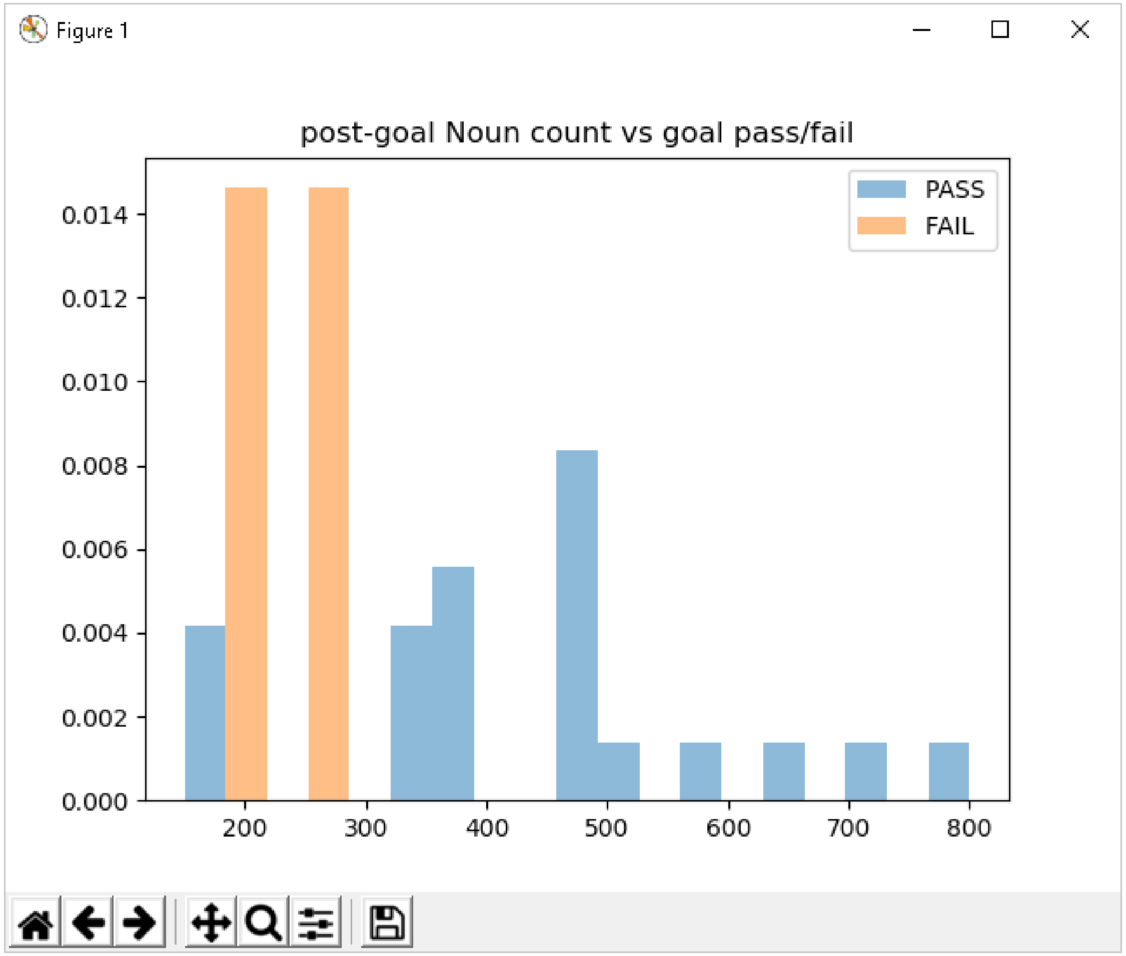
Goal/Objective word length lemmas



Normalised word length



Count of nouns



TF-IDF

- Without a lengthy explanation, in short this weights every word in the SoA to find the words that distinguish a particular document.
- It's not a complex method, but it ends with large arrays of words with numeric representations of their 'importance'

```
X.toarray()[1:5]
          ['ability', 'able',      'abreast',...., 'accept']
array([[0.        , 0.0556693 , 0.        , ..., 0.  ],
       [0.        , 0.01846199, 0.02645575, ..., 0.  ],
       [0.        , 0.        , 0.        , ..., 0.  ],
       [0.        , 0.        , 0.        , ..., 0.  ]])
```

Why did we do that?

- To find ‘dimensions’ that may help classify documents (you don’t have to be sure they will, the model works that out)
- Don’t select related dimensions – e.g. number of paragraphs and document length
- This is where SMEs help – ASIC people who know advice documents and their failings

Machine Learning at last

- Feed the array into ML modelling system..choose multiple parameters with terms like:
  - RandomForestClassifier
  - Accuracy based scoring
  - K_folds (I used 5)
  - N_splits = 2
  - Grid Search of n_estimators [10-300] & max_depths of [30 – infinity]

What did I learn?

- An untrained model:

```
[0.91666667 0.91666667]
```

Split in two sets

```
[1.  0.8 1.  0.8 1. ]
```

Split in five sets

- After fitting and training with holdouts, these were ID'd as the most important dimensions – note that none of my 'extracted features' like noun_counts made it here:

```
[(0.05617283950617285, 'respect'), (0.04444444444444446, 'wish'), (
44444446, 'timeframe'), (0.04444444444444446, 'performing'), (0.044
446, 'option'), (0.03395061728395062, 'indicative'), (0.0327160493827
```

Model results after fitting and training

Using those keywords, on this particular run, I was able to attain 100% precision recall and accuracy – which sounds impressive until you realise there were only two sets of 12 documents, and only two FAILS in the whole 24 documents.

```
Precision: 1.0 / Recall: 1.0/ Accuracy: 1.0
---------------
```

And after a grid_search, which runs repeated tests – the key score here is .916667, which is the fraction 11/12, which means it consistently mislabelled one SoA each run

	mean_fit_time	std_fit_time	mean_score_time	std_score_time	split1_test_score	mean_test_score	std_test_score	rank_test_score
0	0.024982	0.011990	0.002998	0.000000	0.916667	0.916667	0.0	1
1	0.135404	0.002498	0.015989	0.002998	0.916667	0.916667	0.0	1
2	0.296791	0.010992	0.031978	0.007994	0.916667	0.916667	0.0	1
3	0.014989	0.004997	0.002498	0.000500	0.916667	0.916667	0.0	1
4	0.134905	0.022984	0.012491	0.000500	0.916667	0.916667	0.0	1

In summary

- On small dataset, model results don't provide any real result
- What this exercise did teach us:
  - A blend of old and new techniques reaps rewards
  - Involvement from ASIC in classifying datasets for each specific compliance risk is best choice
  - Getting familiar with the data is crucial – analyse, graph, plot
  - We need more data!

Summary

- I could have spent the four weeks generating 4-500 sample SoAs, but that would have given me nothing to show you today.
- The main takeaway is that, if the data exists, code development is not an obstacle

Get in with the machines now!