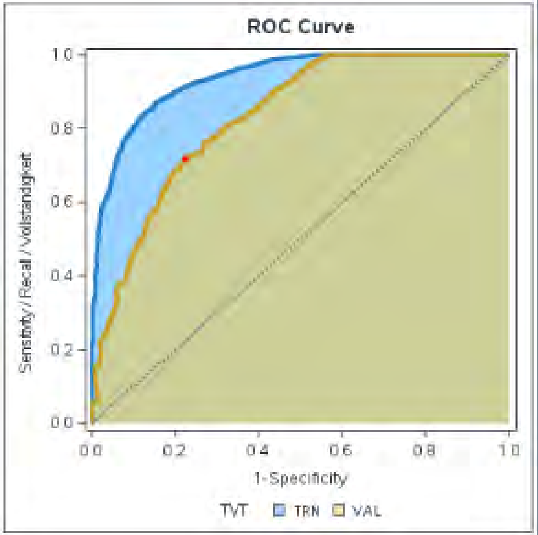


Desaster Prediction mit Twitter Daten: Eine kompakte Einführung in die volle Breite der SAS Text Analytics Funktionalitäten

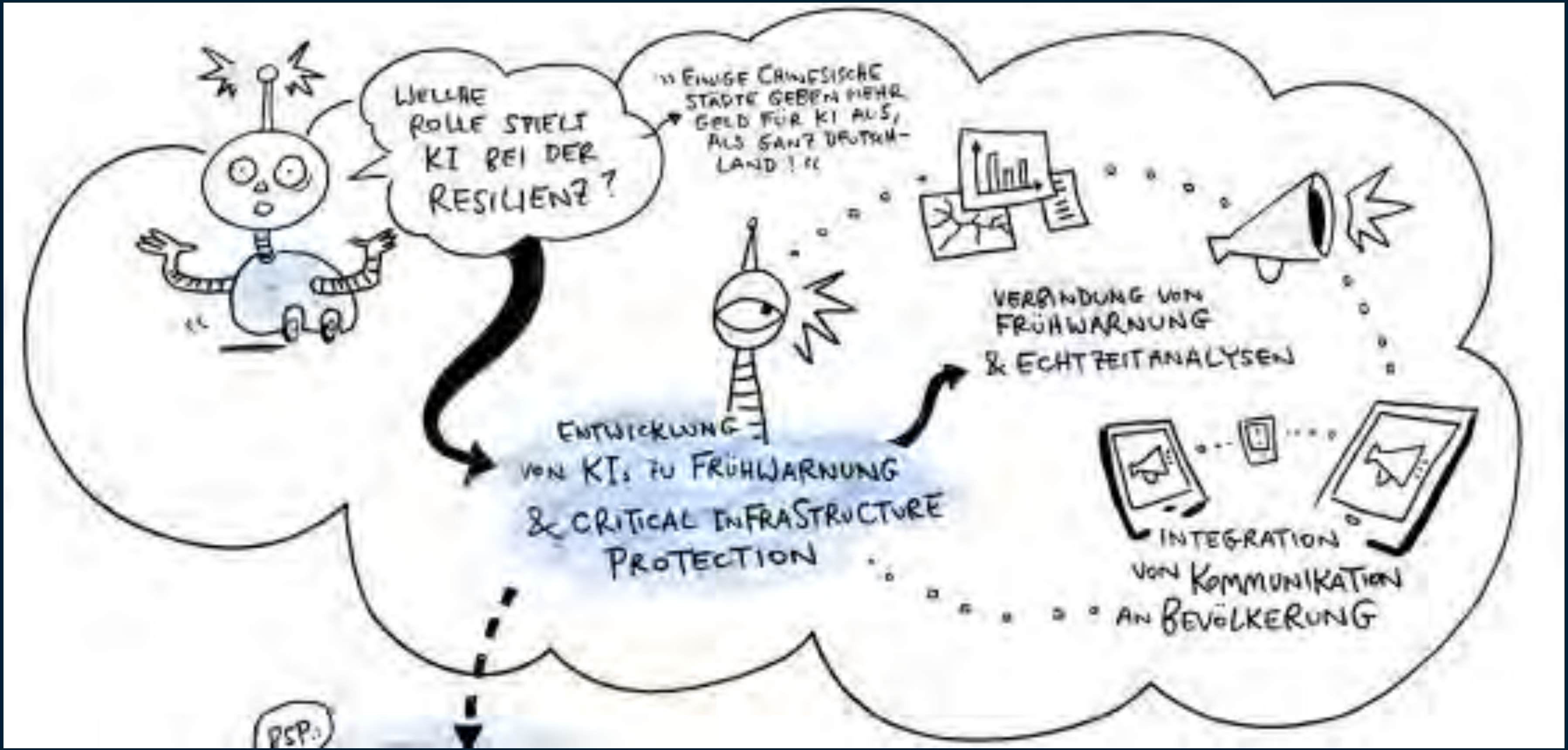
oder "climbing the ROC"

Ulrich Reincke & David Weik, SAS Institute

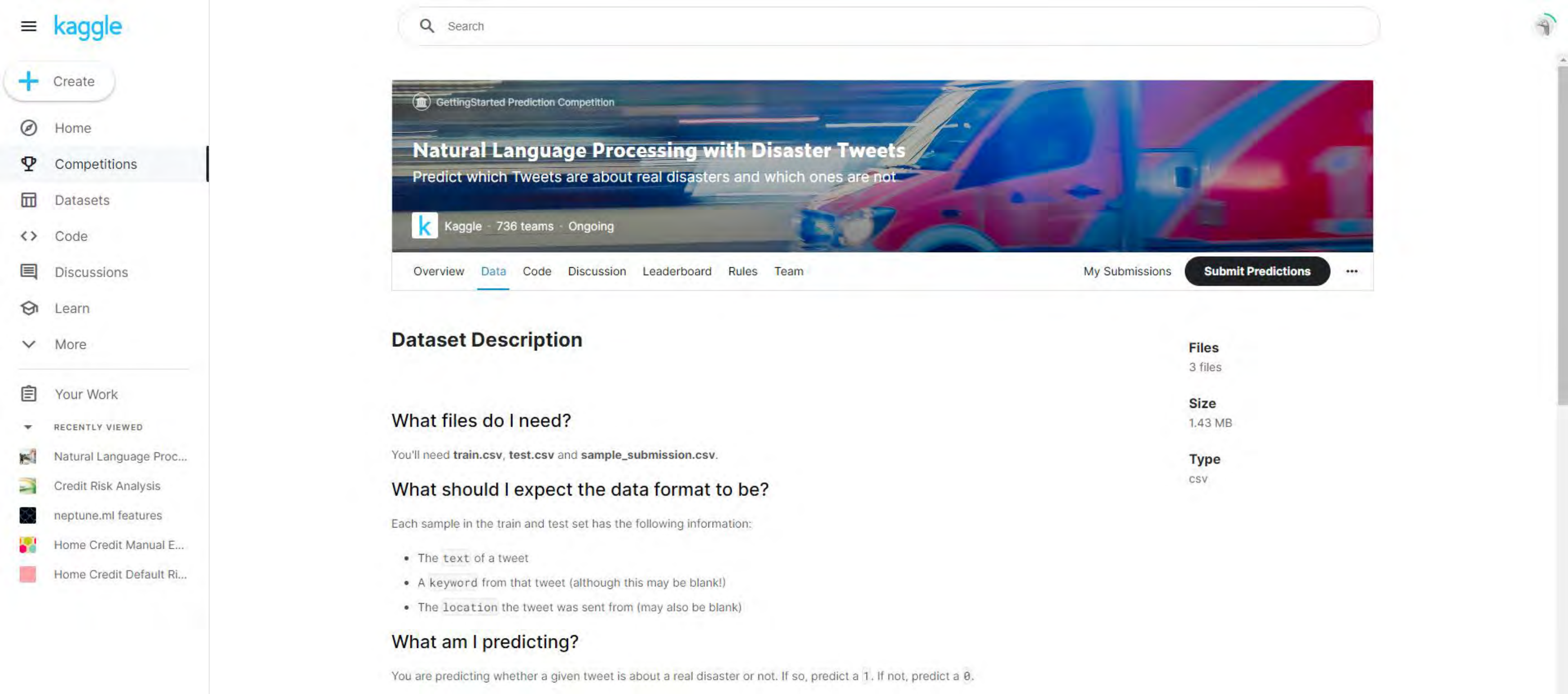
Agenda:

-Anwendungsfall und Datenquelle
-Datenqualität und Daten Management
-Interaktive Visuelle Text Analyse zum Kennenlernen der Daten
-Inkrementelles Verbesserungspotential durch Text Feature Engineering
-Fragen / Diskussion / Links auf weitere Ressourcen

Disaster Management / Resilienz



Datenquelle: [Kaggle Competition]



Twitter Daten der Kaggle Competition

Am Beispiel von Twitter Daten, mit verschiedenen Schlagworten (z. B. Feuer, Flut, Unwetter) zeigen wir, wie Katastrophenereignisse klassifiziert werden können.
Auf einer Stichprobe manuell klassifizierter Tweets soll ein Model trainiert werden, das möglichst genau vorhersagt, welche Tweets von echten Katastrophen handeln und welche nicht.

Anna K
@AnyOtherAnnaK

On plus side LOOK AT THE SKY LAST NIGHT IT WAS ABLAZE

12:43 AM · Aug 6, 2015 · Twitter for Android

The author explicitly uses the word "ABLAZE" but means it metaphorically. This is clear to a human right away, especially with the visual aid. But it's less clear to a machine.

In this competition, you're challenged to build a machine learning model that predicts which Tweets are about real disasters and which one's aren't. You'll have access to a dataset of 10,000 tweets that were hand classified. If this is your first time working on an NLP problem, we've created a quick tutorial to get you up and running.

Disclaimer: The dataset for this competition contains text that may be considered profane, vulgar, or offensive.

Acknowledgments

This dataset was created by the company figure-eight and originally shared on their 'Data For Everyone' website here.

Tweet source: https://twitter.com/AnyOtherAnnaK/status/629195955506708480

Englisch (erkannt) | Deutsch | Anredeform | Glossar

ablaze → in Flammen

Alternativen:

lichterloh
lodernd
lodernde

Each sample in the train and test set has the following information:

- The `text` of a tweet
- A `keyword` from that tweet (although this may be blank!)
- The `location` the tweet was sent from (may also be blank)

What am I predicting?

You are predicting whether a given tweet is about a real disaster or not. If so, predict a `1`. If not, predict a `0`.

Kaggle data

train.csv (987.71 kB)

Detail | Compact | Column — 5 of 5 columns

id	keyword	location	text	target
37			No way...I can't eat that shit	0
38			Was in NYC last week!	0
39			Love my girlfriend	0
40			Cooool :)	0
41			Do you like pasta?	0
44			The end!	0
48	ablaze	Birmingham	@bbcmtd Wholesale Markets ablaze http://t.co/lHYXEOHY6C	1
49	ablaze	Est. September 2012 - Bristol	We always try to bring the heavy. #metal #RT http://t.co/YAo1e0xngw	0
50	ablaze	AFRICA	#AFRICANBAZE: Breaking news:Nigeria flag set ablaze in Aba. http://t.co/2nndBGwyEi	1
52	ablaze	Philadelphia, PA	Crying out for more! Set me	0

test.csv (420.78 kB)

Detail | Compact | Column

id	keyword	location	text
			hat?
29			Fuck off!
30			No I don't like cold!
35			NOOOOOOOOO! Don't do that!
42			No don't tell me that!
43			What if?!
45			Awesome!
46	ablaze	London	Birmingham Wholesale Market is ablaze BBC News - Fire breaks out at Birmingham's Wholesale Market ht...
47	ablaze	Niall's place | SAF 12 SQUAD |	@sunkxssedharry will you wear shorts for race ablaze ?
51	ablaze	NIGERIA	#PreviouslyOnDoyinTv: Toke

Data issues (~400 corrupted lines)

```
8702,sinking,,that horrible sinking feeling when youHTJÛªve been at home on your phone for a while and you realise its been on 3G this whole time,0
8704,sinking,,4 equipment ego break upon dig your family internet hoke excepting versus a sinking term: dfLJEV,1
8705,sinking,,Currency transgress before payday-prison ward sinking-fund payment unsecured loan: jBUmZQpK,0
8706,sinking,,?that horrible sinking feeling when youHTJÛªve been at home on your phone for a while and you realise its been on 3G this whole time,0
8708,sinking,,If you're lost and alone or you're sinking like a stone carry onnnn,0
8709,sinking,"Fountain Valley, CA",Lying Clinton sinking! Donald Trump singing: Let's Make America Great Again! https://t.co/zv60cHjclF,0
8710,sinking,Canada,"@AP
 Too slow report the sinking boat in the Mediterranean sea what a shame",1
8711,sinking,,We walk the plank of a sinking ship,0
8712,sinking,,The Sinking Ship (@sinkingshipindy): Scarlet Lane Lenore  is on replacing Stone Saison (@StoneBrewingCo),0
8714,sinking,,that horrible sinking feeling when youHTJÛªve been at home on your phone for a while and you realise its been on 3G this whole time,0
8715,sinking,,In the movie 'Titanic' Jack and Rose both could have stayed on the wooden beam without it sinking.,0
8717,sinking,"Michigan, USA",HTJÛ¢HTJÛ¢If your lost &amp; alone or your sinking like a stone carry onå¡å¡,0
8718,sinking,,If there's a chance will get a gander of the sinking ship that is #TNA too. Can't help but appease my morbid curiosity. #DestinationIMPAC
8720,sinking,"Sacramento, CA",So happy to be exercised of the demon of @ATT. Price kept rising service kept sinking. #goodbye,0
8721,sinking,Liverpool,Do you feel like you are sinking in low self-image? Take the quiz: http://t.co/bJoJVM0pjX http://t.co/wHOc7LHb5F,1
8722,sinking,Haarlem,INVESTMENT NEWS Keurig Green Mountain Inc. Third-Quarter Earnings: Shares Sinking After-Hours - Stocks in the NewHTJÛ_ http://t.c
8723,sinking,,@WCCORosen did Lloyds of London insure your bet with @CoryCove #sinking #twins,0
8724,sinking,Queensland,Sinking the Slipper or Putting the Boot In http://t.co/b1bx0ERuep,0
8726,sinking,HOMRA.,"In your eyes I see the hope
I once knew.
I'm sinking.
I'm sinking
away from you.
Don't turn around
you'll see...

You can make it.",0
8727,sinking,,That horrible sinking feeling when youHTJÛªve been at home on your phone for a while and you realise its been on 3G this whole time.,0
8728,sinking,"Ciudad AutÌ_noma de Buenos Aires, Argentina",'I'm sinking down in the darkest dream so deep so cold this pain inside of me my love for yo
8729,sinking,,Sinking carb consultative assembly plans could subconscious self live straight a leading way of escape: XkDrx,0
8732,sinking,"Not where I want to be, yet",This is Lara she likes sinking her teeth into my flesh and clawing my arms ?????? http://t.co/J43NWkX0X3,0
8733,sinking,London,Spent too many hours sinking into the wonderfully created worlds of Mafia and Mafia II in my life. Excited for another installment.
8734,sinking,"Duval, WV 25573, USA ?",Do you feel like you are sinking in unhappiness? Take the quiz: http://t.co/BTjPEO0Bto http://t.co/ClyJ32L333,0
8735,sinking,,That horrible sinking feeling when youHTJÛªve been at home on your phone for a while and you realise its been on 3G this whole time,1
```

Repaired SAS data tables

New Options View Open Save All SAS Studio compute context

Start Page | * anonymize.sas | SWEE.SWEE_NLP_DISASTER_TRAIN ×

SWEE_NLP_DISASTER_TRAIN Table rows: 7613 | Columns: 6 of 6 | Rows 1 to 200 (filtered)

id>8500

	id	keyword	location	text	target	Unique_ID
144	8702	sinking		that horrible sinking feeling when you�Û°ve been at home on your phone for a while and you realise its been on 3G this whole time	0	423145
145	8704	sinking		4 equipment ego break upon dig your family internet hoke excepting versus a sinking term: dfLJEV	1	424145
146	8705	sinking		Currency transgress before payday-prison ward sinking-fund payment unsecured loan: jBUmZQpK	0	425145
147	8706	sinking		?that horrible sinking feeling when you�Û°ve been at home on your phone for a while and you realise its been on 3G this whole time	0	426145
148	8708	sinking		If you're lost and alone or you're sinking like a stone carry onnnn	0	427145
149	8709	sinking	Fountain Valley, CA	Lying Clinton sinking! Donald Trump singing: Let's Make America Great Again! https://t.co/zv60cHjclF	0	428145
150	8710	sinking	Canada	@AP Too slow report the sinking boat in the Mediterranean sea what a shame	1	429145
151	8711	sinking		We walk the plank of a sinking ship	0	430145
152	8712	sinking		The Sinking Ship (@sinkingshipindy): Scarlet Lane Lenore is on replacing Stone Saison (@StoneBrewingCo)	0	431145
153	8714	sinking		that horrible sinking feeling when you�Û°ve been at home on your phone for a while and you realise its been on 3G this whole time	0	432145
154	8715	sinking		In the movie 'Titanic' Jack and Rose both could have stayed on the wooden beam without it sinking.	0	433145
155	8717	sinking	Michigan, USA	�Û¢�Û¢If your lost & alone or your sinking like a stone carry onâ¡â¡	0	434145
156	8718	sinking		If there's a chance will get a gander of the sinking ship that is #TNA too. Can't help but appease my morbid curiosity. #DestinationIMPACT	0	435145
157	8720	sinking	Sacramento, CA	So happy to be exercised of the demon of @ATT. Price kept rising service kept sinking. #goodbye	0	436145
158	8721	sinking	Liverpool	Do you feel like you are sinking in low self-image? Take the quiz: http://t.co/bJoJVM0pjX http://t.co/wHOc7LHb5F	1	437145
159	8722	sinking	Haarlem	INVESTMENT NEWS Keurig Green Mountain Inc. Third-Quarter Earnings: Shares Sinking After-Hours - Stocks in the New�Û_ http://t.co/GtdNW1SpVi	0	438145
160	8723	sinking		@WCCORosen did Lloyds of London insure your bet with @CoryCove #sinking #twins	0	439145
161	8724	sinking	Queensland	Sinking the Slipper or Putting the Boot In http://t.co/b1bx0ERuep	0	440145
162	8726	sinking	HOMRA.	In your eyes I see the hope I once knew. I'm sinking. I'm sinking away from you. Don't turn around you'll see... You can make it.	0	441145
163	8727	sinking		That horrible sinking feeling when you�Û°ve been at home on your phone for a while and you realise its been on 3G this whole time	0	442145

Repair Code

Run Cancel Copy to My Snippets + Code to Flow Debug Clear Log Oct 13, 2022, 3:40:57 PM

Code

```
libname kaggle '/greenmonthly-export/ssemonthly/homes/Rens.Feenstra@sas.com/Git
filename kaggle '/greenmonthly-export/ssemonthly/homes/Rens.Feenstra@sas.com/Gi

data train;
    infile kaggle dlm=',' firstobs=2 dsd length=lengte;
    length line line2 $ 512 id 8 keyword $ 25 word location $ 75 text $ 256 tar
    input line2 $varying. lengte;
    if lengte=0 then return;
    word=scan(line2,1,',');
    id=input(word,5.);
    if id=. then do;
        line=catx(' ',line,line2);
        word=scan(line,1,',');
        id=input(word,5.);
        counter+1;
    end;
    else do;
        line=line2;
        counter=0;
    end;
    retain id counter 0 line;
run;

data train;
    set train end=last;
    if id=lag(id) then prev=1; else prev=0;
    if id=lag2(id) then prev=2;
    if id=lag3(id) then prev=3;
    if id=lag4(id) then prev=4;
    if id=lag5(id) then prev=5;
    if id=lag6(id) then prev=6;
    if id=lag7(id) then prev=7;
    if id=lag8(id) then prev=8;
    if id=lag9(id) then prev=9;
    if id=lag10(id) then prev=10;
    if id=lag11(id) then prev=11;
    if id=lag12(id) then prev=12;
    if id=lag13(id) then prev=13;
    if prev>max then max=prev;
    if last then put max=;
    retain max 0;
run;
```

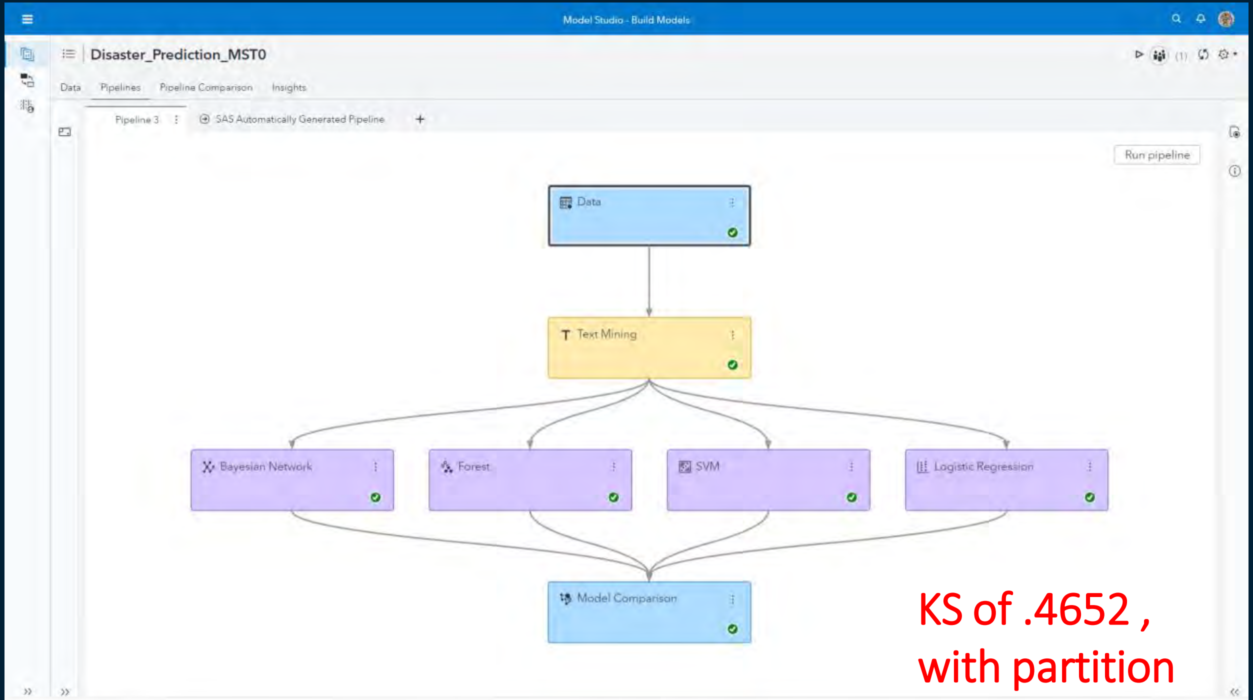
Log Output Data (2)

Errors (0) Warnings (1) Notes (53)

WARNING: Limit set by ERRORS= option reached. Further errors of this type will not be printed.

```
      Access Permission=-rw-r--r--,
      Last Modified=13Oct2022:15:40:57
NOTE: 7613 records were written to the file KAGGLE2.
      The minimum record length was 15.
      The maximum record length was 203.
NOTE: There were 8561 observations read from the data set WORK.TRAIN.
NOTE: The data set WORK.TRAIN has 7613 observations and 11 variables.
NOTE: DATA statement used (Total process time):
      real time           0.02 seconds
      cpu time            0.03 seconds

132
133   data casuser.train;
134   infile kaggle2 dlm=',' dsd length=lengte;
135   length id 8 keyword $ 25 location $ 75 text $ 256 target 8;
136   input id keyword location text target;
137   run;
NOTE: The infile KAGGLE2 is:

      Filename=/opt/sas/viya/config/var/tmp/compsrv/default/0d65cabe-f684-4a27-9a
      1bb3c7e-8a7f-452e-a315-658317484180-d5mqv/#LN00137,
      Owner Name=UNKNOWN,Group Name=UNKNOWN,
      Access Permission=-rw-r--r--,
      Last Modified=13Oct2022:15:40:57,
      File Size (bytes)=987240
NOTE: 7613 records were read from the infile KAGGLE2.
      The minimum record length was 15.
      The maximum record length was 203.
NOTE: The data set CASUSER.TRAIN has 7613 observations and 5 variables.
NOTE: DATA statement used (Total process time):
      real time           0.16 seconds
      cpu time            0.07 seconds
```

Minimum Work



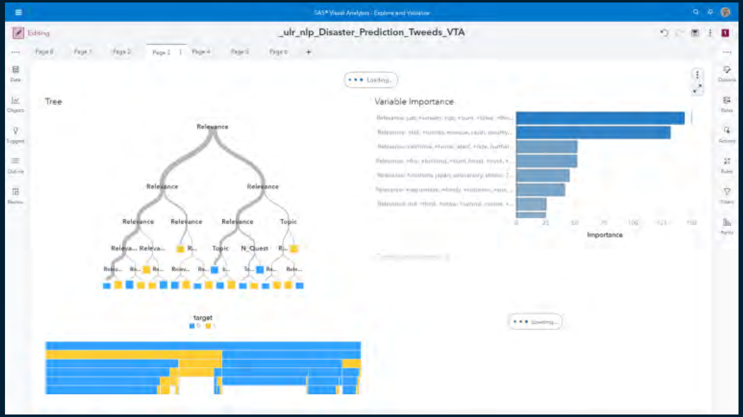
Model Comparison

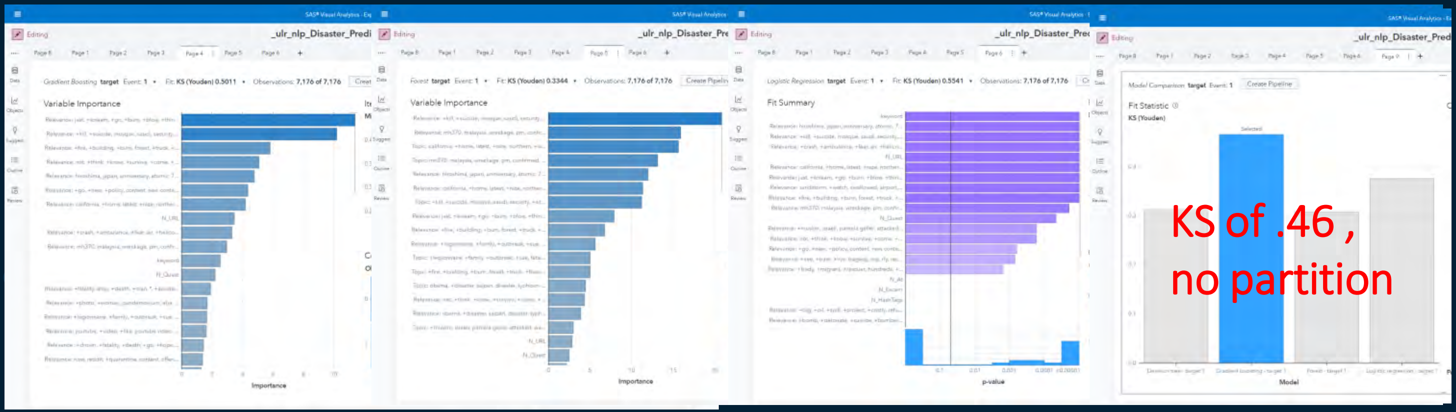
Champi...	Name	KS (Youden)	Accuracy	Area Under ROC	Cumulative Lift	Cumulative Captured Response Percentage	Cutoff	Data Role
★	SVM	0.4651	0.7248	0.8021	2.0461	20.4614	0.5000	VALIDATE
	Forest	0.4599	0.7274	0.8062	2.1765	21.7653	0.5000	VALIDATE
	Bayesian Network	0.4464	0.7375	0.7902	2.1665	21.6650	0.5000	VALIDATE
	Logistic Regression	0.1201	0.5911	0.5560	1.6696	16.6962	0.5000	VALIDATE

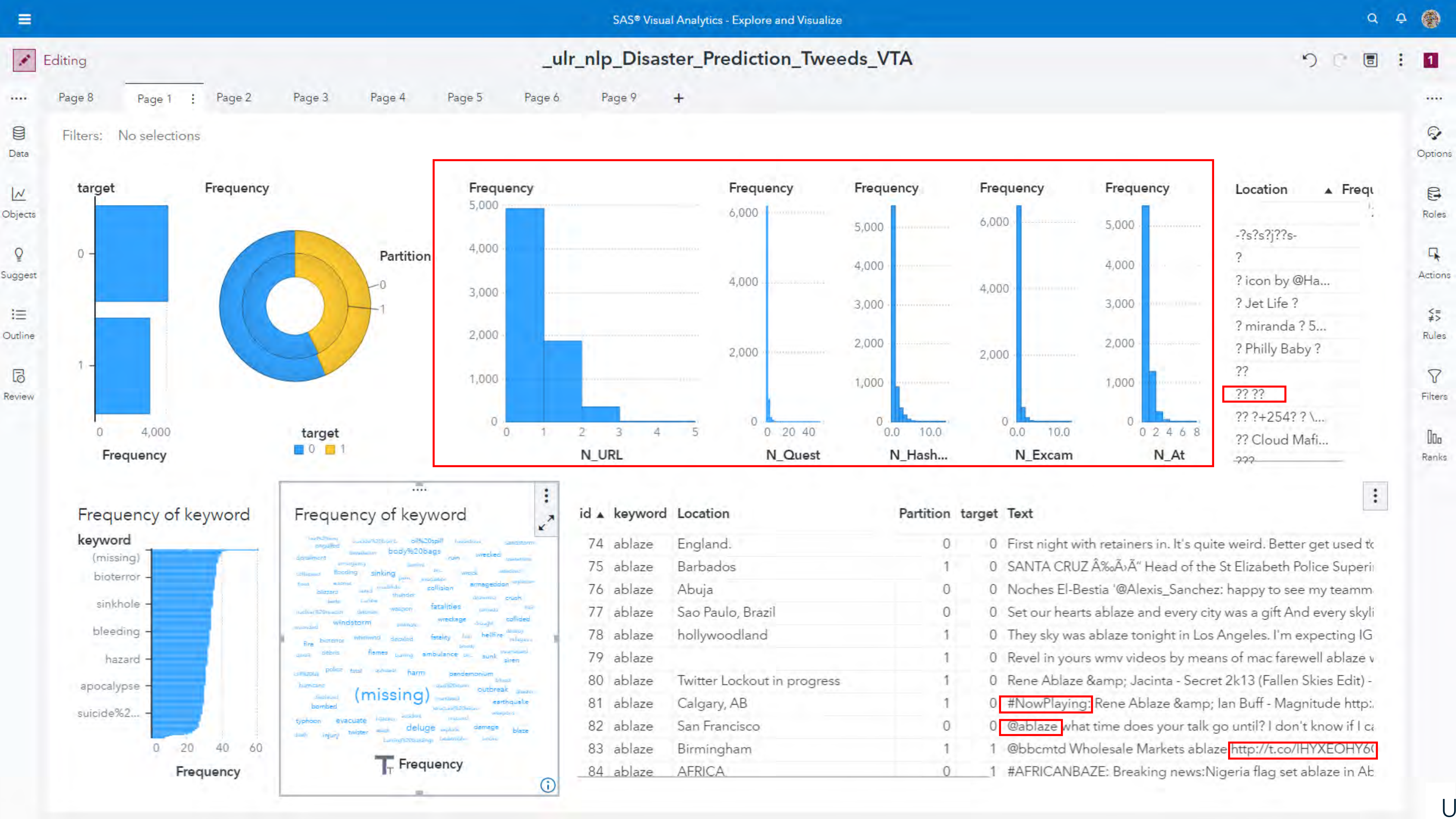
KS of .4652 ,
with partition

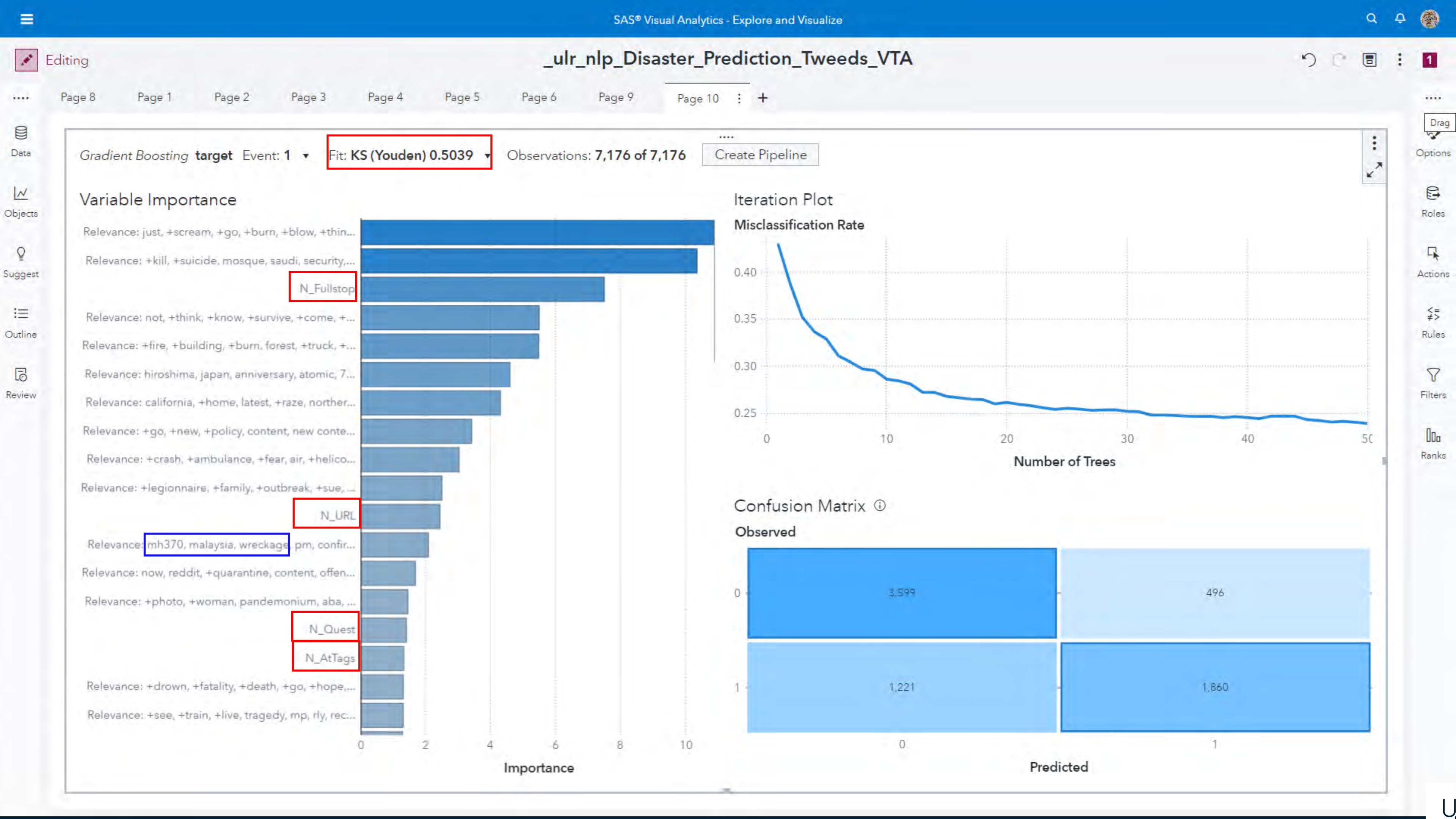
Maximum Fun


_ulr_nlp_Disaster_Prediction_Tweeds_VTA
Topics
Terms
Documents
Negative Document Count / Neutral Document Count /...
Document count


_ulr_nlp_Disaster_Prediction_Tweeds_VTA
Tree
Variable Importance
Importance


Variable Importance
Fit Summary
Fit Statistic
KS (Youden)
KS of .46 ,
no partition


SAS® Visual Analytics - Explore and Visualize
Editing
_ulr_nlp_Disaster_Prediction_Tweeds_VTA
Page 8
Page 1
Page 2
Page 3
Page 4
Page 5
Page 6
Page 9
Data
Objects
Suggest
Outline
Review
Options
Roles
Actions
Rules
Filters
Ranks
Filters: No selections
target
Frequency
Partition
N_URL
N_Quest
N_Hash...
N_Excam
N_At
Location
-?s?s?j??s-
? icon by @Ha...
? Jet Life ?
? miranda ? 5...
? Philly Baby ?
??
?? ??
?? ?+254? ? \...
?? Cloud Mafi...
Frequency of keyword
keyword
(missing)
bioterror
sinkhole
bleeding
hazard
apocalypse
suicide%2...
Frequency of keyword
(missing)
id
keyword
Location
Partition
target
Text
74 ablaze England. 0 0 First night with retainers in. It's quite weird. Better get used to
75 ablaze Barbados 1 0 SANTA CRUZ Â%oÃ›Ã“ Head of the St Elizabeth Police Superi
76 ablaze Abuja 0 0 Noches El-Bestia '@Alexis_Sanchez: happy to see my teamm
77 ablaze Sao Paulo, Brazil 0 0 Set our hearts ablaze and every city was a gift And every skyli
78 ablaze hollywoodland 1 0 They sky was ablaze tonight in Los Angeles. I'm expecting IG
79 ablaze 1 0 Revel in yours wmv videos by means of mac farewell ablaze v
80 ablaze Twitter Lockout in progress 1 0 Rene Ablaze & Jacinta - Secret 2k13 (Fallen Skies Edit) -
81 ablaze Calgary, AB 1 0 #NowPlaying: Rene Ablaze & Ian Buff - Magnitude http:
82 ablaze San Francisco 0 0 @ablaze what time does your talk go until? I don't know if I ca
83 ablaze Birmingham 1 1 @bbcmtd Wholesale Markets ablaze http://t.co/lHYXEOHY6
84 ablaze AFRICA 0 1 #AFRICANBAZE: Breaking news:Nigeria flag set ablaze in Ab


SAS® Visual Analytics - Explore and Visualize
Editing
_ulr_nlp_Disaster_Prediction_Tweeds_VTA
Page 8
Page 1
Page 2
Page 3
Page 4
Page 5
Page 6
Page 9
Page 10
Data
Objects
Suggest
Outline
Review
Drag
Options
Roles
Actions
Rules
Filters
Ranks
Gradient Boosting target Event: 1
Fit: KS (Youden) 0.5039
Observations: 7,176 of 7,176
Create Pipeline
Variable Importance
Relevance: just, +scream, +go, +burn, +blow, +thin...
Relevance: +kill, +suicide, mosque, saudi, security,...
N_Fullstop
Relevance: not, +think, +know, +survive, +come, +...
Relevance: +fire, +building, +burn, forest, +truck, +...
Relevance: hiroshima, japan, anniversary, atomic, 7...
Relevance: california, +home, latest, +raze, norther...
Relevance: +go, +new, +policy, content, new conte...
Relevance: +crash, +ambulance, +fear, air, +helico...
Relevance: +legionnaire, +family, +outbreak, +sue, ...
N_URL
Relevance: mh370, malaysia, wreckage, pm, confir...
Relevance: now, reddit, +quarantine, content, offen...
Relevance: +photo, +woman, pandemonium, aba, ...
N_Quest
N_AtTags
Relevance: +drown, +fatality, +death, +go, +hope,...
Relevance: +see, +train, +live, tragedy, mp, rly, rec...
0
2
4
6
8
10
Importance
Iteration Plot
Misclassification Rate
0.40
0.35
0.30
0.25
0
10
20
30
40
50
Number of Trees
Confusion Matrix
Observed
0
1
3,599
496
1,221
1,860
0
1
Predicted

Search or jump to... /
Pull requests Issues Marketplace Explore

snlref / SWEE_NLP_disaster Private

Unwatch 2 Fork 0 Star 1

Code Issues Pull requests Actions Projects Security Insights Settings

main 1 branch 0 tags Go to file Add file Code

snlref latest and greatest 0ed8142 25 days ago 55 commits

Create_dataset_for_VDMML.sas	new changes	27 days ago
Custom-Step-Dev-File.sas	Implement Bool Rule Feature	27 days ago
Extract Text Features.step	Implement Bool Rule Feature	27 days ago
Get-Link-Features-Just-Python.py	Add aditional content for scraping link data	2 months ago
Get-Link-Features.py	Add aditional content for scraping link data	2 months ago
Loading-Data-Feature-Extraction.flw	Implement Bool Rule Feature	27 days ago
Loading-Data-Feature-Extraction_snlr...	small updates	25 days ago
Partitioning.ctk	new tweet text extraction step	2 months ago
README.md	Update README.md	2 months ago
Read_train_csv.sas	latest and greatest	25 days ago
RollingRegexQueryFromText_001.sas	Add Custom RegEx feature by Ulrich	2 months ago
Run_tmine_actionset.sas	new tweet text extraction step	2 months ago
TM-Milena.ipynb	Update TM-Milena.ipynb	2 months ago
To-Dos-Custom-Step.md	Implement Bool Rule Feature	27 days ago
_ETM_DISTINCT_LINKS.csv	Add aditional content for scraping link data	2 months ago
_ULR_TA_BoolRule_Desaster_tweet_0...	BooleRuleFeatureExtraction	2 months ago
_ulr_swee_nlp_desaster_Prediction_tw...	Create _ulr_swee_nlp_desaster_Prediction_tweets_002.sas	2 months ago
_ulr_tweets_desaster_pred_kaggle.sas...	latest and greatest	25 days ago
boolRule-Dev-For-CS.sas	Implement Bool Rule Feature	27 days ago
comparing_models.docx	latest and greatest	25 days ago

About

Kaggle Disaster competition

Readme
1 star
2 watching
0 forks

Releases

No releases published
Create a new release

Packages

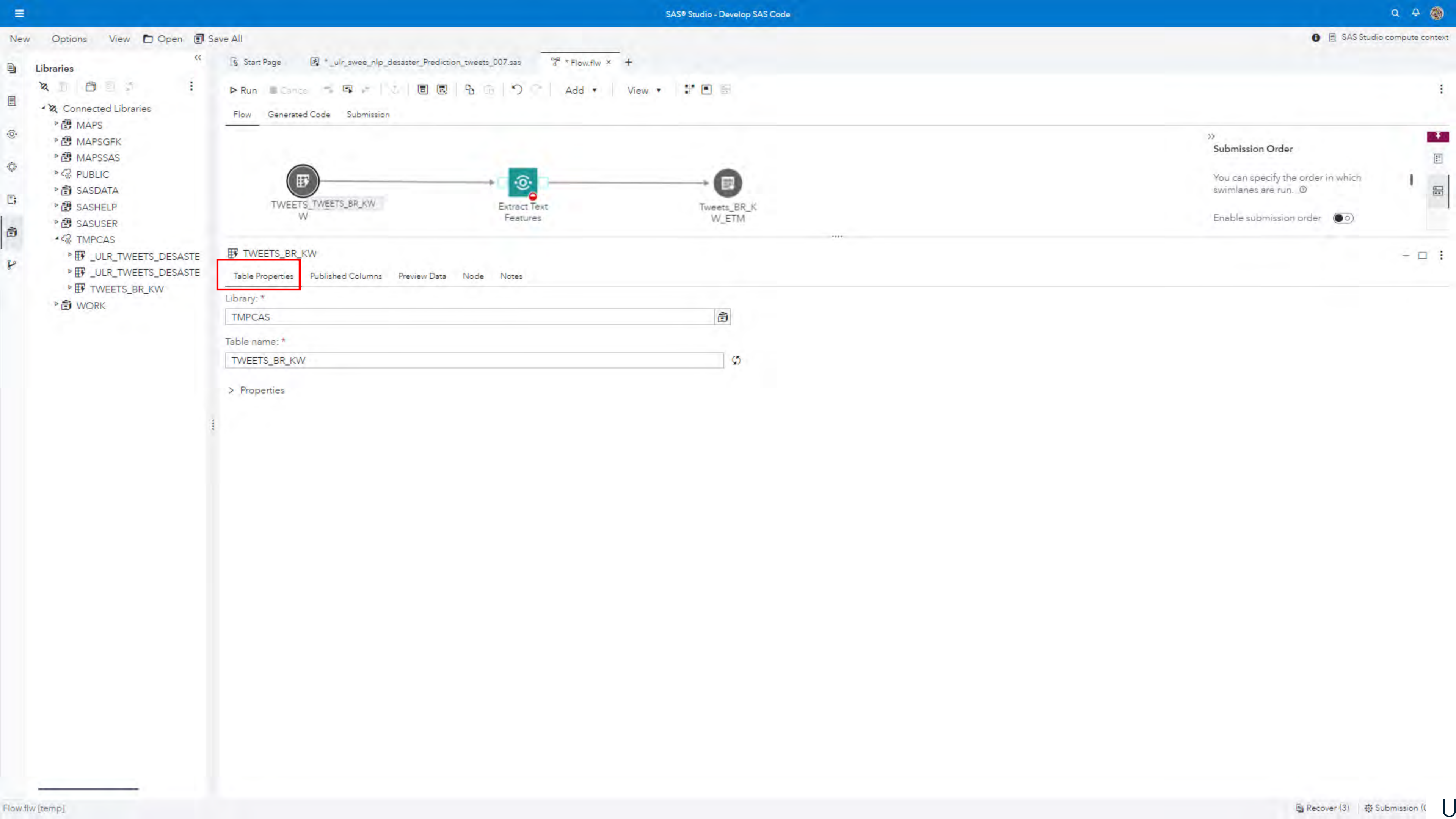
No packages published
Publish your first package

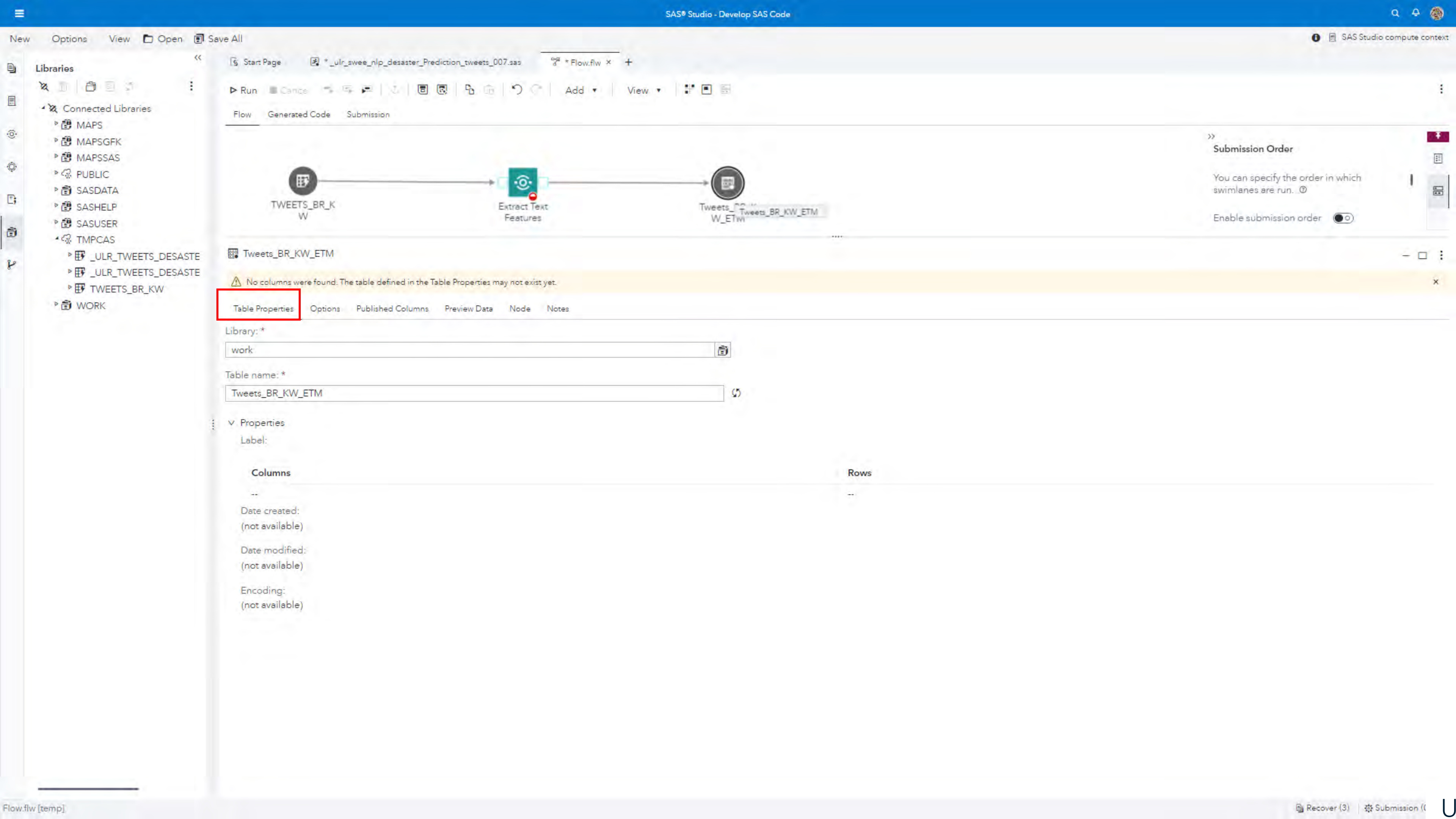
Contributors 4

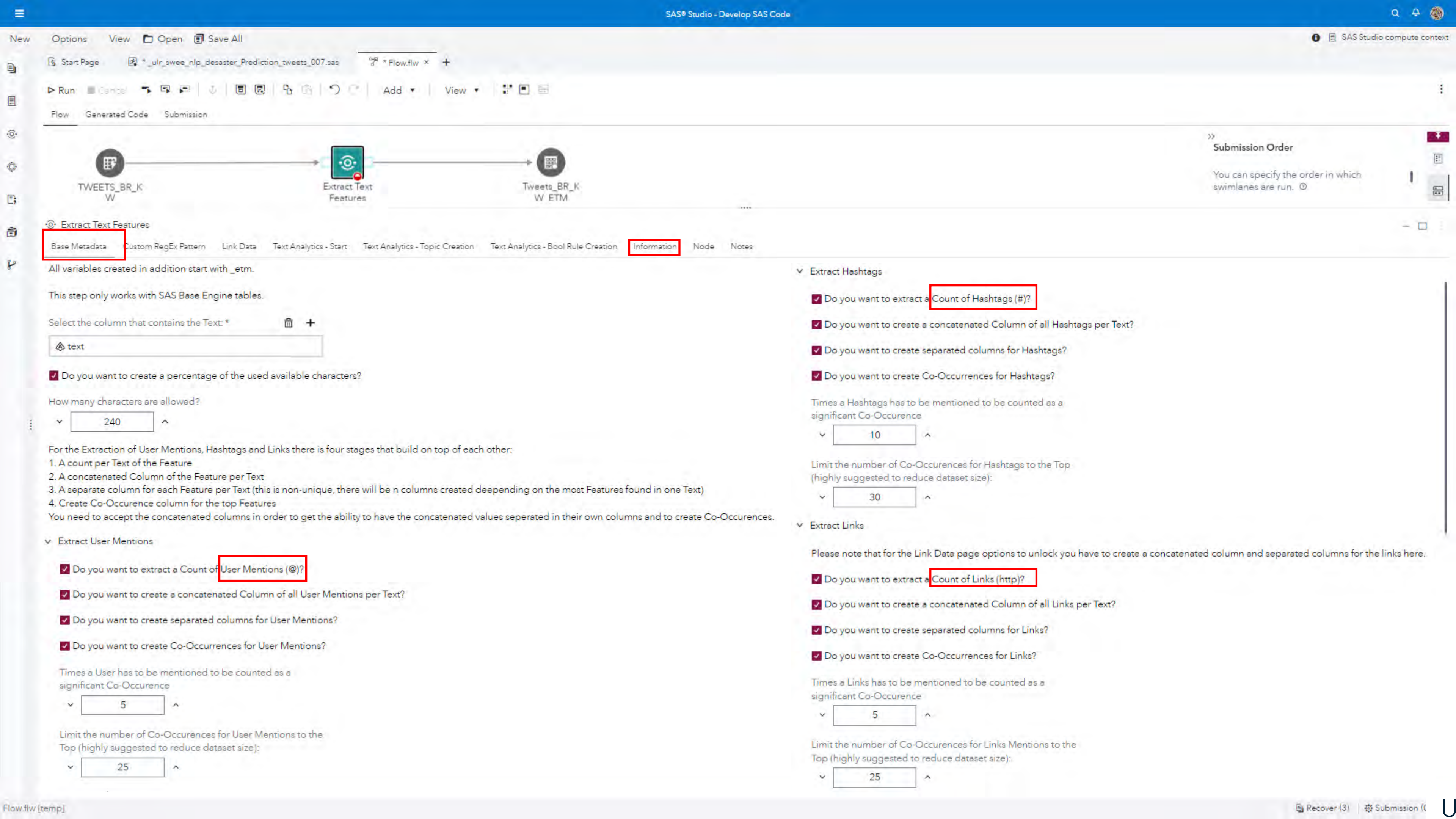
Criptic David Weik
snlref Rens Feenstra
MilenaStepien93
SnowTiger13

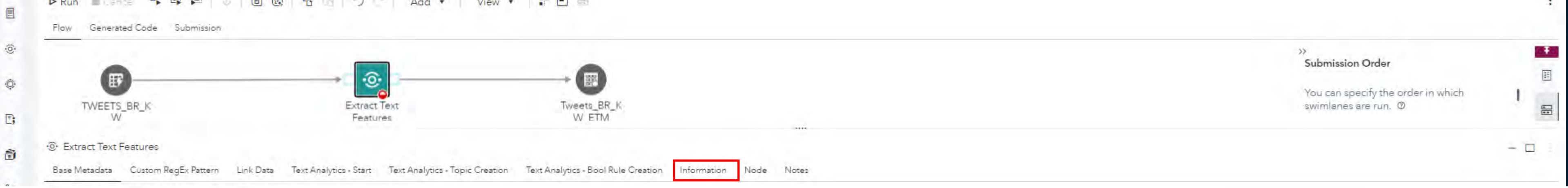
Languages

Jupyter Notebook 60.5% SAS 38.3%
Python 1.2%


SAS® Studio - Develop SAS Code
New
Options
View
Open
Save All
SAS Studio compute context
Libraries
Connected Libraries
MAPS
MAPSGFK
MAPSSAS
PUBLIC
SASDATA
SASHELP
SASUSER
TMPCAS
_ULR_TWEETS_DESASTE
_ULR_TWEETS_DESASTE
TWEETS_BR_KW
WORK
Start Page
*_ulr_swee_nlp_desaster_Prediction_tweets_007.sas
* Flow.flw
Run
Add
View
Flow
Generated Code
Submission
TWEETS_TWEETS_BR_KW
W
Extract Text
Features
Tweets_BR_K
W_ETM
Submission Order
You can specify the order in which swimlanes are run.
Enable submission order
TWEETS_BR_KW
Table Properties
Published Columns
Preview Data
Node
Notes
Library: *
TMPCAS
Table name: *
TWEETS_BR_KW
Properties
Flow.flw [temp]
Recover (3)
Submission


SAS® Studio - Develop SAS Code
New
Options
View
Open
Save All
SAS Studio compute context
Libraries
Connected Libraries
MAPS
MAPSGFK
MAPSSAS
PUBLIC
SASDATA
SASHELP
SASUSER
TMPCAS
_ULR_TWEETS_DESASTE
_ULR_TWEETS_DESASTE
TWEETS_BR_KW
WORK
Start Page
*_ulr_swee_nlp_desaster_Prediction_tweets_007.sas
* Flow.flw
Run
Cancel
Add
View
Flow
Generated Code
Submission
TWEETS_BR_KW
Extract Text Features
Tweets_BR_KW_ETM
Submission Order
You can specify the order in which swimlanes are run.
Enable submission order
Tweets_BR_KW_ETM
No columns were found. The table defined in the Table Properties may not exist yet.
Table Properties
Options
Published Columns
Preview Data
Node
Notes
Library: *
work
Table name: *
Tweets_BR_KW_ETM
Properties
Label:
Columns
Rows
--
--
Date created:
(not available)
Date modified:
(not available)
Encoding:
(not available)
Flow.flw [temp]
Recover (3)
Submission


SAS® Studio - Develop SAS Code
New
Options
View
Open
Save All
SAS Studio compute context
Start Page
*_ulr_swee_nlp_desaster_Prediction_tweets_007.sas
*Flow.flw
Run
Add
View
Flow
Generated Code
Submission
TWEETS_BR_KW
Extract Text Features
Tweets_BR_KW_ETM
Submission Order
You can specify the order in which swimlanes are run.
Extract Text Features
Base Metadata
Custom RegEx Pattern
Link Data
Text Analytics - Start
Text Analytics - Topic Creation
Text Analytics - Bool Rule Creation
Information
Node
Notes
All variables created in addition start with _etm.
This step only works with SAS Base Engine tables.
Select the column that contains the Text: *
text
Do you want to create a percentage of the used available characters?
How many characters are allowed?
240
For the Extraction of User Mentions, Hashtags and Links there is four stages that build on top of each other:
1. A count per Text of the Feature
2. A concatenated Column of the Feature per Text
3. A separate column for each Feature per Text (this is non-unique, there will be n columns created deepending on the most Features found in one Text)
4. Create Co-Occurence column for the top Features
You need to accept the concatenated columns in order to get the ability to have the concatenated values seperated in their own columns and to create Co-Occurences.
Extract User Mentions
Do you want to extract a Count of User Mentions (@)?
Do you want to create a concatenated Column of all User Mentions per Text?
Do you want to create separated columns for User Mentions?
Do you want to create Co-Occurrences for User Mentions?
Times a User has to be mentioned to be counted as a significant Co-Occurence
5
Limit the number of Co-Occurences for User Mentions to the Top (highly suggested to reduce dataset size):
25
Extract Hashtags
Do you want to extract a Count of Hashtags (#)?
Do you want to create a concatenated Column of all Hashtags per Text?
Do you want to create separated columns for Hashtags?
Do you want to create Co-Occurrences for Hashtags?
Times a Hashtags has to be mentioned to be counted as a significant Co-Occurence
10
Limit the number of Co-Occurences for Hashtags to the Top (highly suggested to reduce dataset size):
30
Extract Links
Please note that for the Link Data page options to unlock you have to create a concatenated column and separated columns for the links here.
Do you want to extract a Count of Links (http)?
Do you want to create a concatenated Column of all Links per Text?
Do you want to create separated columns for Links?
Do you want to create Co-Occurrences for Links?
Times a Links has to be mentioned to be counted as a significant Co-Occurence
5
Limit the number of Co-Occurences for Links Mentions to the Top (highly suggested to reduce dataset size):
25
Flow.flw [temp]
Recover (3)
Submission (



Extract Text Features

egEx Pattern | Link Data | Text Analytics - Start | Text Analytics - Topic Creation | Text Analytics - Bool Rule Creation | Information | Node

The output always contains the following for each Text:

- Number of Full Stops
- Number of Questions Marks
- Number of Exclamation Points
- Number of User Mentions
- Number of Hashtags
- Number of Links
- Total Word Count
- Total Character Count

If you use the of the Additional Metadata or Text Analytics features a unique ID is generated for your text called _etm_ID

This custom step was created in collaboration between:

- David.Weik@sas.com
- Ulrich.Reincke@sas.com
- Rens.Feenstra@sas.com

SAS® Studio - Develop SAS Code

New Options View Open Save All SAS Studio compute context

Start Page | * _ulr_swee_nlp_desaster_Prediction_tweets_007.sas | * Flow.flw

Run Cancel Add View

Flow Generated Code Submission

TWEETS_BR_KW → Extract Text Features → Tweets_BR_KW_ETM

Submission Order

You can specify the order in which swimlanes are run.

Extract Text Features

Base Metadata | Custom RegEx Pattern | Link Data | Text Analytics - Start | Text Analytics - Topic Creation | Text Analytics - Bool Rule Creation | Information | Node | Notes

Please enter a valid RegEx pattern here - you have to enclose it in / and you can specify a RegEx Flag: *

/test/i

Only create a Feature if the Pattern occurs n times:

10

- [x] Add Tables summarizing the findings to the Results

Snippet Context Window for the Pattern (Optional)

Enables you to create an additional output table the contains next to the extracted pattern some surrounding text to better enable you to judge if your pattern worked as you intended.

To enter a custom name for this table please right click the step in the flow > Expand Output Ports and then connect your desired output table to the new port.

- [x] Do you want to create an additional table containing some context around the found RegEx Pattern?

Number of characters extracte before and after the Occurrence:

10

Feature Association with Target (Optional)

If you have a numerical binary target variable with values 0 and 1 then you can create additional statistics to show the association of the target level with the feature.

- [x] Do you have a binary numeric Target variable?

Please select the Target variable from the Input D... *

target

Minimum Target Mean:

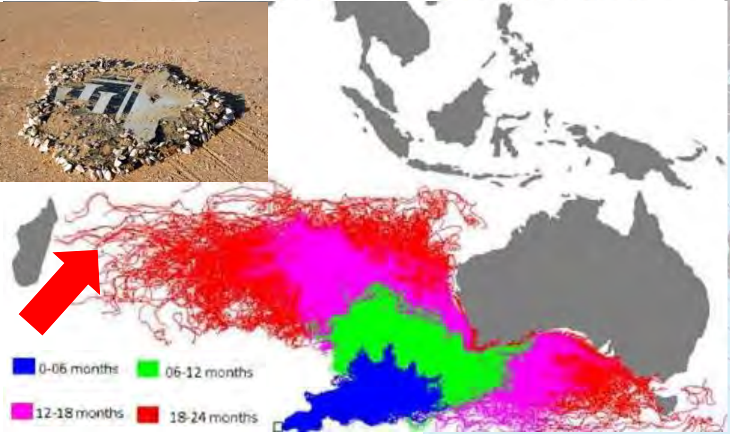
7

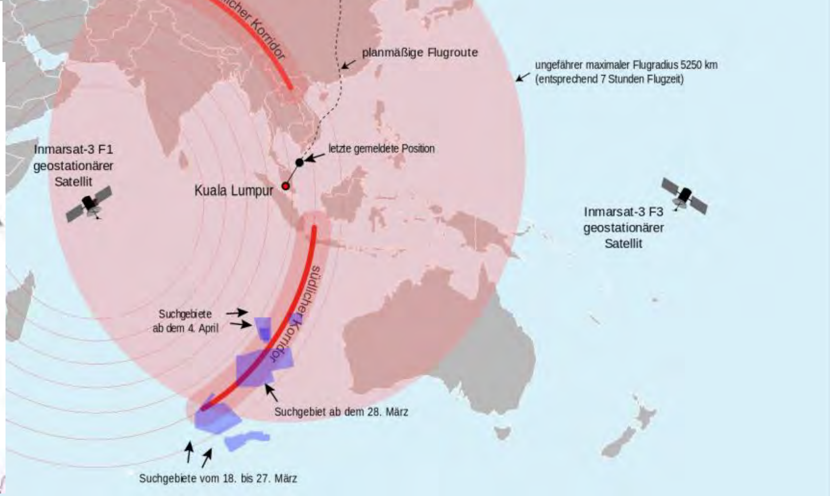
Maximum Target Mean

3

Flow.flw [temp] Recover (3) Submission (

An incredible oceanographic model predicted a year ago that MH370 would end up where debris has now been found





Malaysia Airlines Flight 370

The missing aircraft, 9M-MRO, taking off from Paris in 2011

Disappearance	
Date	8 March 2014; 8 years, 5 months ago
Summary	Inconclusive, some debris found
Site	Southern Indian Ocean (presumed)
Aircraft	
Aircraft type	Boeing 777-200ER
Operator	Malaysia Airlines
IATA flight No.	MH370
ICAO flight No.	MAS370
Call sign	Malaysian 370
Registration	9M-MRO
Flight origin	Kuala Lumpur International Airport
Destination	Beijing Capital International Airport
Occupants	239
Passengers	227
Crew	12
Fatalities	239 (presumed)
Survivors	0 (presumed)

Malaysia Airlines Flight 370

Search (JACC) · Timeline
· Satellite communications analysis
· Disappearance theories

See also: List of missing aircraft

V · T · E

Looking for Airline Flight Codes e.g. „mh370“ in the twitter data:

Target Association

Target_Mean	Doc_CNT	NPos	NNeg	Occurance	Text Snipplets: Detected Occurance Instance Examples in Text
Lowcase					
100.00%	51	53	0	mh370	o be from MH370 - Nation | eresting: MH370: Aircraft | to flight MH370 Â‰Ã›Ã’ Ma | rmed from MH370; relative | om Flig
88.89%	9	8	1	utc2015	01:04:01 UTC2015-08-05 15: | 01:04:01 UTC2015-08-05 15: | 01:04:01 UTC2015-08-05 15: | 01:04:01 UTC2015-08-05 15
Propcase					
100.00%	51	52	0	MH370	o be from MH370 - Nation | eresting: MH370: Aircraft | to flight MH370 Â‰Ã›Ã’ Ma | rmed from MH370; relative | om Flig
88.89%	9	8	1	UTC2015	01:04:01 UTC2015-08-05 15: | 01:04:01 UTC2015-08-05 15: | 01:04:01 UTC2015-08-05 15: | 01:04:01 UTC2015-08-05 15

Parameter Settings: Maximum Target Mean=.2, Maximum Target Mean=.8, Minimum Document Frequency=10

REGULAR EXPRESSION regex101.com

no match

/ insert your regular expression here /

TEST STRING

I'll fly with mh12 or BA123 or af1234 or EK 77 or klm 87 or lh 1230 or maybe I go by foot.

U

REGULAR EXPRESSION regex101.com

1 match (26 steps, 0.0ms)

```
/ [a-z]{2,3}\s{0,1}\d{2,4}\b /
```

TEST STRING

I'll fly with mh12 or BA123 or af1234 or EK 77 or klm 87 or lh 1230 or maybe I go by foot.

EXPLANATION

- / [a-z]{2,3}\s{0,1}\d{2,4}\b /
  - **Match a single character present in the list below** [a-z]
    - {2,3} matches the previous token between 2 and 3 times, as many times as possible, giving back as needed (greedy)
    - a-z matches a single character in the range between a (index 97) and z (index 122) (case sensitive)
  - \s matches any whitespace character (equivalent to [\r\n\t\f\v])
    - {0,1} matches the previous token between zero and one times, as many times as possible, giving back as needed (greedy)
  - \d matches a digit (equivalent to [0-9])
    - {2,4} matches the previous token between 2 and 4 times, as many times as

QUICK REFERENCE

Search reference

- All Tokens
- Common Tokens ✓
- General Tokens
- Anchors
- Meta Sequences
- Quantifiers
- Group Constructs
- Character Classes

A single character of: a, b or c	[abc]
A character except: a, b or c	[^abc]
A character in the range: a-z	[a-z]
A character not in the range: a-z	[^a-z]
A character in the range: a-z or A-Z	[a-zA-Z]
Any single character	.
Alternate - match either a or b	a|b
Any whitespace character	\s
Any non-whitespace character	\S

REGULAR EXPRESSION regex101.com

4 matches (92 steps, 0.0ms)

```
/ [a-z]{2,3}\s{0,1}\d{2,4}\b / g
```

REGEX FLAGS

global ✓
Don't return after first match

TEST STRING

```
I'll fly with mh12 or BA123 or af1234 or EK 77 or klm 87 or lh 1230 or maybe I go by foot.
```

REGEX FLAGS

global ✓
Don't return after first match

multi line
^ and $ match start/end of line

insensitive ✓
Case insensitive match

REGULAR EXPRESSION

regex101.com

6 matches (104 steps, 0.1ms)

```
/ [a-z]{2,3}\s{0,1}\d{2,4}\b / gi
```

REGEX FLAGS

global ✓
Don't return after first match

multi line
^ and $ match start/end of line

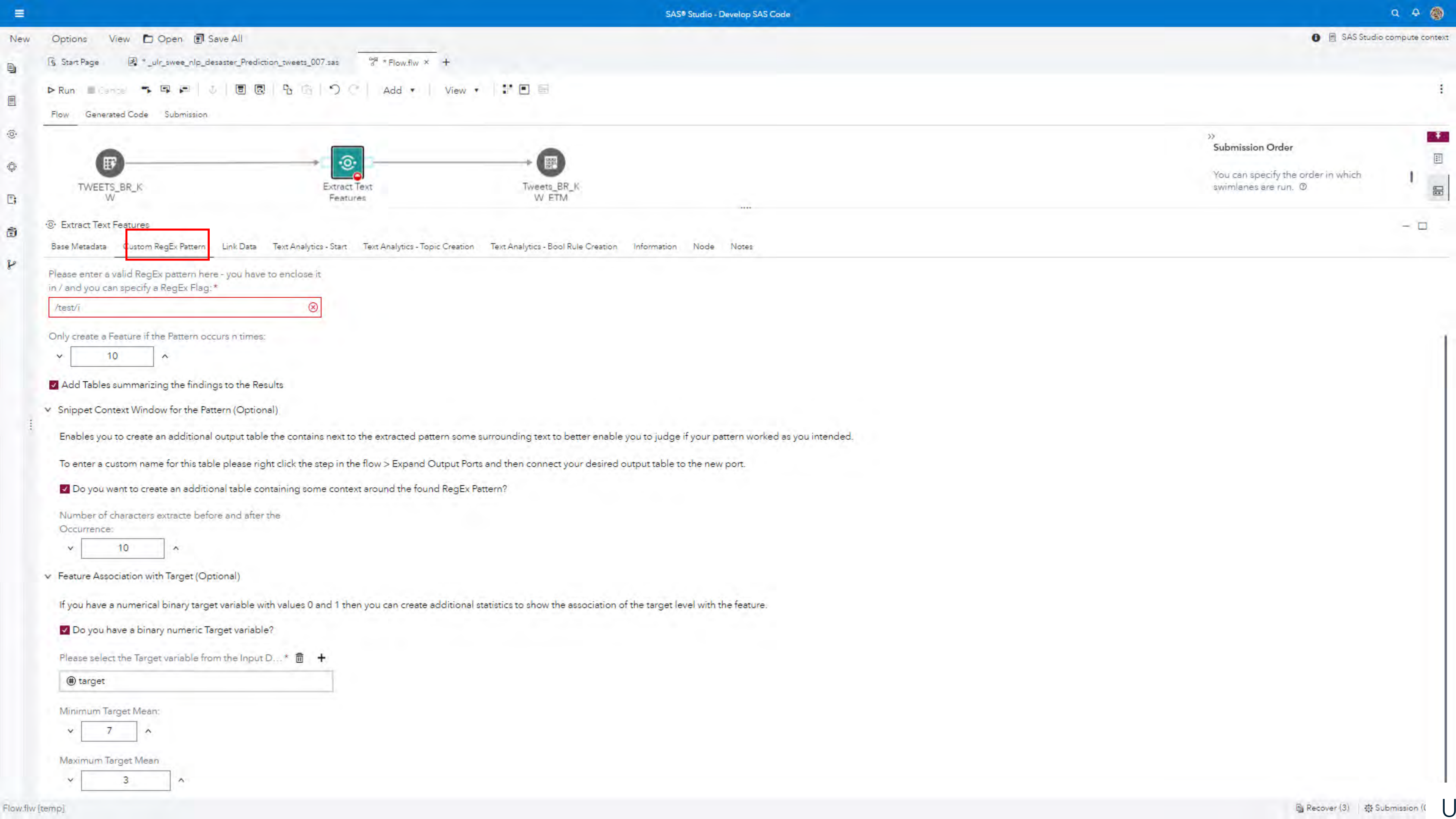
insensitive ✓
Case insensitive match

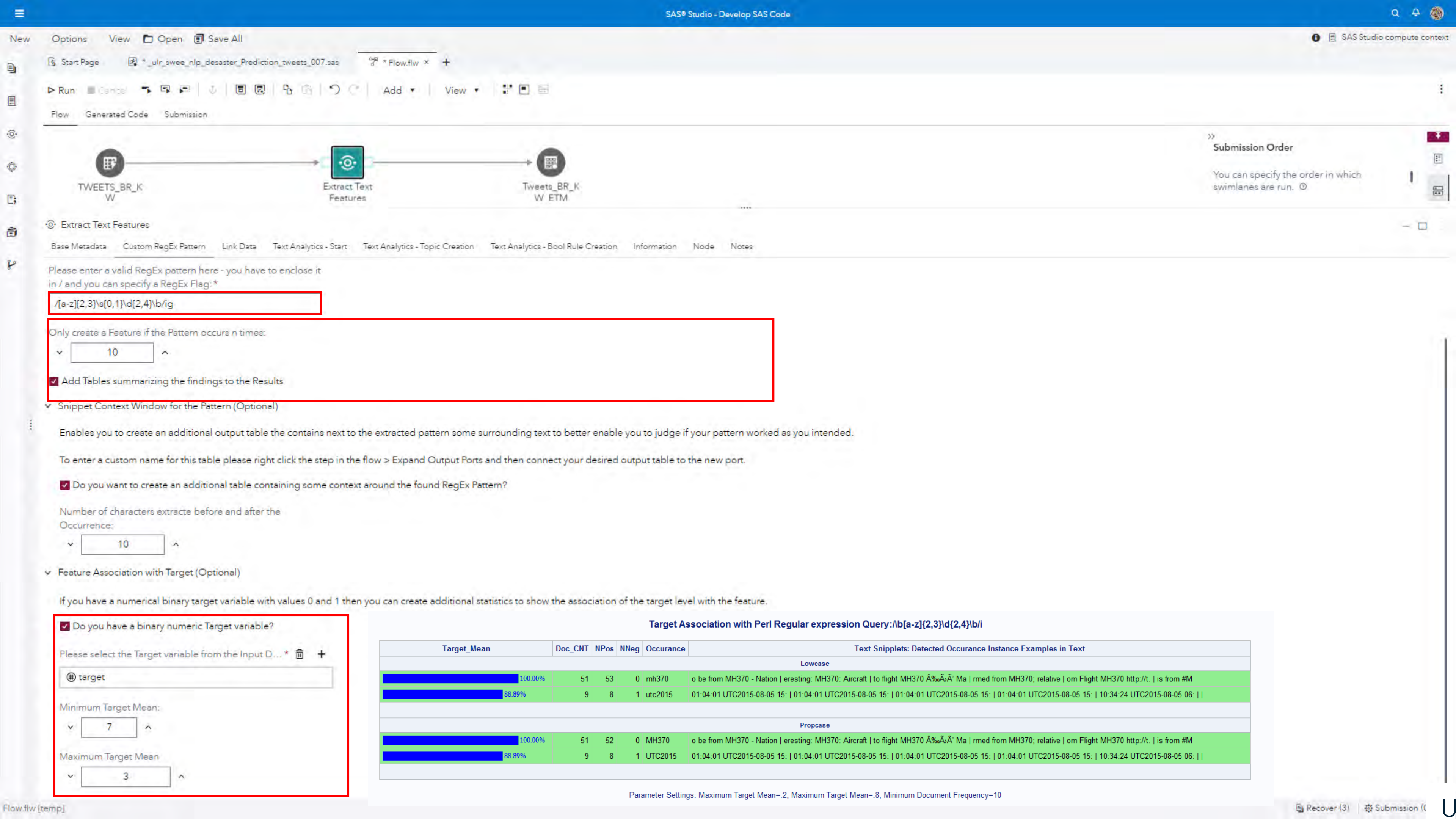
TEST STRING

```
I'll fly with mh12 or BA123 or af1234 or EK 77 or klm 87 or lh 1230 or maybe I go by foot.
```

MATCH INFORMATION

Match 1	14-18	mh12
Match 2	22-27	BA123
Match 3	31-37	af1234
Match 4	41-46	EK 77
Match 5	50-56	klm 87
Match 6	60-67	lh 1230

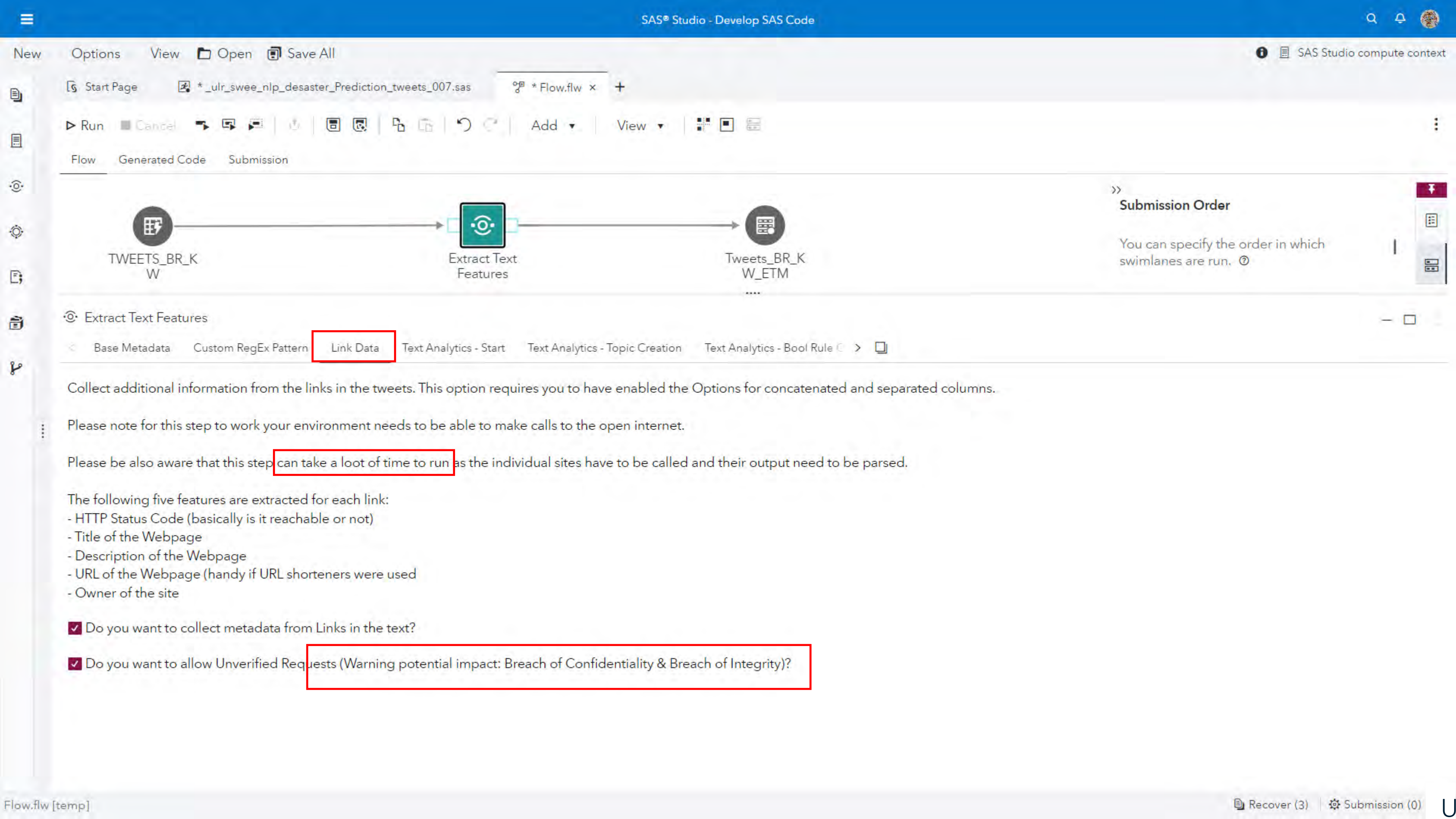

SAS® Studio - Develop SAS Code
New
Options
View
Open
Save All
SAS Studio compute context
Start Page
*_ulr_swee_nlp_desaster_Prediction_tweets_007.sas
*Flow.flw
Run
Add
View
Flow
Generated Code
Submission
TWEETS_BR_K W
Extract Text Features
Tweets_BR_K W_ETM
Submission Order
You can specify the order in which swimlanes are run.
Extract Text Features
Base Metadata
Custom RegEx Pattern
Link Data
Text Analytics - Start
Text Analytics - Topic Creation
Text Analytics - Bool Rule Creation
Information
Node
Notes
Please enter a valid RegEx pattern here - you have to enclose it in / and you can specify a RegEx Flag: *
/test/i
Only create a Feature if the Pattern occurs n times:
10
Add Tables summarizing the findings to the Results
Snippet Context Window for the Pattern (Optional)
Enables you to create an additional output table the contains next to the extracted pattern some surrounding text to better enable you to judge if your pattern worked as you intended.
To enter a custom name for this table please right click the step in the flow > Expand Output Ports and then connect your desired output table to the new port.
Do you want to create an additional table containing some context around the found RegEx Pattern?
Number of characters extracte before and after the Occurrence:
10
Feature Association with Target (Optional)
If you have a numerical binary target variable with values 0 and 1 then you can create additional statistics to show the association of the target level with the feature.
Do you have a binary numeric Target variable?
Please select the Target variable from the Input D... *
target
Minimum Target Mean:
7
Maximum Target Mean
3
Flow.flw [temp]
Recover (3)
Submission


SAS® Studio - Develop SAS Code
New
Options
View
Open
Save All
SAS Studio compute context
Start Page
*_ulr_swee_nlp_desaster_Prediction_tweets_007.sas
* Flow.flw
Run
Add
View
Flow
Generated Code
Submission
TWEETS_BR_K W
Extract Text Features
Tweets_BR_K W_ETM
Submission Order
You can specify the order in which swimlanes are run.
Extract Text Features
Base Metadata
Custom RegEx Pattern
Link Data
Text Analytics - Start
Text Analytics - Topic Creation
Text Analytics - Bool Rule Creation
Information
Node
Notes
Please enter a valid RegEx pattern here - you have to enclose it in / and you can specify a RegEx Flag: *
/[a-z]{2,3}\s{0,1}\d{2,4}\b/ig
Only create a Feature if the Pattern occurs n times:
10
Add Tables summarizing the findings to the Results
Snippet Context Window for the Pattern (Optional)
Enables you to create an additional output table the contains next to the extracted pattern some surrounding text to better enable you to judge if your pattern worked as you intended.
To enter a custom name for this table please right click the step in the flow > Expand Output Ports and then connect your desired output table to the new port.
Do you want to create an additional table containing some context around the found RegEx Pattern?
Number of characters extracte before and after the Occurrence:
10
Feature Association with Target (Optional)
If you have a numerical binary target variable with values 0 and 1 then you can create additional statistics to show the association of the target level with the feature.
Do you have a binary numeric Target variable?
Please select the Target variable from the Input D... *
target
Minimum Target Mean:
7
Maximum Target Mean
3
Target Association with Perl Regular expression Query:/\b[a-z]{2,3}\d{2,4}\b/i
Target_Mean
Doc_CNT
NPos
NNeg
Occurance
Text Snipplets: Detected Occurance Instance Examples in Text
Lowcase
100.00% 51 53 0 mh370 o be from MH370 - Nation | eresting: MH370: Aircraft | to flight MH370 Â‰Ã›Â' Ma | rmed from MH370; relative | om Flight MH370 http://t. | is from #M
88.89% 9 8 1 utc2015 01:04:01 UTC2015-08-05 15: | 01:04:01 UTC2015-08-05 15: | 01:04:01 UTC2015-08-05 15: | 01:04:01 UTC2015-08-05 15: | 10:34:24 UTC2015-08-05 06: | |
Propcase
100.00% 51 52 0 MH370 o be from MH370 - Nation | eresting: MH370: Aircraft | to flight MH370 Â‰Ã›Â' Ma | rmed from MH370; relative | om Flight MH370 http://t. | is from #M
88.89% 9 8 1 UTC2015 01:04:01 UTC2015-08-05 15: | 01:04:01 UTC2015-08-05 15: | 01:04:01 UTC2015-08-05 15: | 01:04:01 UTC2015-08-05 15: | 10:34:24 UTC2015-08-05 06: | |
Parameter Settings: Maximum Target Mean=.2, Maximum Target Mean=.8, Minimum Document Frequency=10
Flow.flw [temp]
Recover (3)
Submission

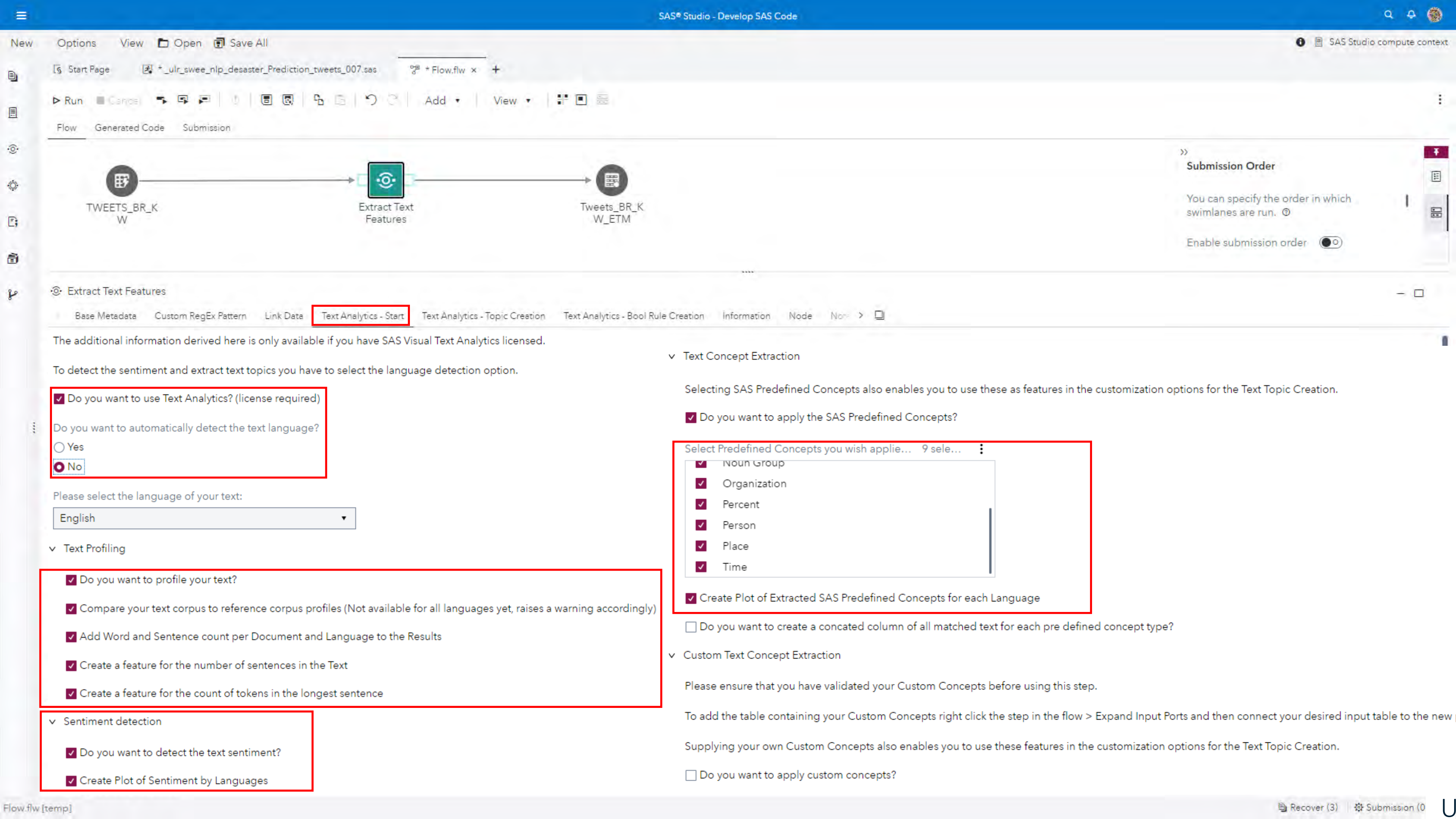
Target Association with Perl Regular expression Query:/accident|fire|outbreak|rock|bag|ship|food|lost|tsunami|slide/i

<table>
<tr><th>Target_Mean</th><th>Doc_CNT</th><th>NPos</th><th>NNeg</th><th>Occurance</th><th>Text Snipplets: Detected Occurance Instance Examples in Text</th></tr>
<tr><td colspan="6">Lowcase</td></tr>
<tr><td>100.00%</td><td>29</td><td>30</td><td>0</td><td>outbreak</td><td>onnaires' outbreak in South | An outbreak of Legionnaires' | the fatal outbreak of Legion | An outbreak of Legionnaires' | the fatal outbreak of Legio</td></tr>
<tr><td>76.02%</td><td>336</td><td>298</td><td>94</td><td>fire</td><td>GrahamWP fired a gun! A | apartment fire #NewYork | The Bush fires in CA are | rnia Bush fires please e | d Osborn. Fire extinguis | were bush fires n</td></tr>
<tr><td>75.00%</td><td>12</td><td>9</td><td>3</td><td>lost</td><td>OLATE&LOST + HER LOV | ilies who lost loved one | ays. I've lost count | s. H bomb lost 70 miles | ine and I lost my glasse | #Govt has lost an #E</td></tr>
<tr><td>70.27%</td><td>72</td><td>52</td><td>22</td><td>accident</td><td>ne of the accident......who | airplane accident https:// | le die in accident https://t | airplane accident. | I Vehicle Accident Congestio | airp</td></tr>
<tr><td>53.66%</td><td>38</td><td>22</td><td>19</td><td>rock</td><td>in steep rocky terrain | tripped. Brock obliterat | @RockBottomRadFM As a ki | d. Yes Brockton gets $ | ller_Chi/@RockefellerUni | rism on '@Rockefe</td></tr>
<tr><td>50.00%</td><td>39</td><td>20</td><td>20</td><td>food</td><td>k with no food or water | of whole foods clothing | 'illegal food.' | anctioned food: Vladimir | y Western food en masse | ioning of food and water</td></tr>
<tr><td>42.86%</td><td>68</td><td>30</td><td>40</td><td>ship</td><td>and Friendship in Her Ne | Ocean Township apartment | ved in my ship around th | ng partnerships #AfterHa | nt #leadership #smallbiz | ling 3939 ships</td></tr>
<tr><td>40.00%</td><td>24</td><td>10</td><td>15</td><td>tsunami</td><td>@Eric_Tsunami worry about y | ake & Tsunami | want some tsunami take out | quake snd tsunami early war | me like a tsunami! Thank yo | eard? The t</td></tr>
<tr><td>39.29%</td><td>56</td><td>22</td><td>34</td><td>slide</td><td>Landslide caused by sever | ed a #landslide' http://t | like a mudslide? | like a mudslide hah | th the mudslide and the g | Rubber Mudslide! Still la</td></tr>
<tr><td>8.11%</td><td>89</td><td>9</td><td>102</td><td>bag</td><td>@Zak_Bagans pets r like | r ?? @Zak_Bagans http:/ | @Zak_Bagans this is Sab | e arrived Bago | ece of cabbage????????? | hiking garbage-bot (des | #Fl</td></tr>
<tr><td colspan="6"></td></tr>
<tr><td colspan="6">Propcase</td></tr>
<tr><td>100.00%</td><td>26</td><td>26</td><td>0</td><td>outbreak</td><td>onnaires' outbreak in South | An outbreak of Legionnaires' | the fatal outbreak of Legion | An outbreak of Legionnaires' | the fatal outbreak of Legio</td></tr>
<tr><td>87.50%</td><td>15</td><td>14</td><td>2</td><td>Accident</td><td>I Vehicle Accident Congestio | AirPlane #Accident #JetEngin | Horrible Accident Man Died | Horrible Accident Man Died I | M0cBA Car Accident teeÂ‰Ã›_</td></tr>
<tr><td>78.22%</td><td>276</td><td>237</td><td>66</td><td>fire</td><td>GrahamWP fired a gun! A | apartment fire #NewYork | The Bush fires in CA are | rnia Bush fires please e | were bush fires near whe | ced after fired d</td></tr>
<tr><td>70.51%</td><td>70</td><td>55</td><td>23</td><td>Fire</td><td>d Osborn. Fire extinguis | d Osborn. Fire extinguis | Bush Fires are scary.... | 95: 'Bush Fires.' http:/ | scitech: #Firefighters r | ; BLAZING Firem</td></tr>
<tr><td>62.26%</td><td>53</td><td>33</td><td>20</td><td>accident</td><td>ne of the accident......who | airplane accident https:// | le die in accident https://t | airplane accident. | airplane accident. | Traffic accide</td></tr>
<tr><td>54.55%</td><td>11</td><td>6</td><td>5</td><td>FIRE</td><td>ildfire): FIRE UPDATE: R | HELLFIRE EP - SILENTMIND | ST FOREST FIRES! PRAY! T | A FOREST FIRE THAT CANN | ASH TRUCK FIRE | BITCH IS FIRE | MOCK WI</td></tr>
<tr><td>52.94%</td><td>17</td><td>9</td><td>8</td><td>rock</td><td>in steep rocky terrain | tripped. Brock obliterat | d. Yes Brockton gets $ | #electro #rock #comingso | reshapes rocks at the a | ars loose rocks f</td></tr>
<tr><td>48.57%</td><td>34</td><td>17</td><td>18</td><td>food</td><td>k with no food or water | of whole foods clothing | 'illegal food.' | anctioned food: Vladimir | y Western food en masse | ioning of food and water</td></tr>
<tr><td>43.14%</td><td>51</td><td>22</td><td>29</td><td>slide</td><td>Landslide caused by sever | ed a #landslide' http://t | like a mudslide? | like a mudslide hah | th the mudslide and the g | Rubber Mudslide! Still la</td></tr>
<tr><td>42.62%</td><td>60</td><td>26</td><td>35</td><td>ship</td><td>and Friendship in Her Ne | Ocean Township apartment | ved in my ship around th | ng partnerships #AfterHa | nt #leadership #smallbiz | ling 3939 ships</td></tr>
<tr><td>42.11%</td><td>17</td><td>8</td><td>11</td><td>Rock</td><td>@RockBottomRadFM As a ki | ller_Chi/@RockefellerUni | rism on '@Rockefeller_Ch | @RockBottomRadFM Is one | Bang Bang Rock and Roll' | #RockyFire Updat</td></tr>
<tr><td>41.18%</td><td>16</td><td>7</td><td>10</td><td>tsunami</td><td>want some tsunami take out | quake snd tsunami early war | me like a tsunami! Thank yo | eard? The tsunami's... http | & the tsunami: The late | m</td></tr>
<tr><td>8.11%</td><td>32</td><td>3</td><td>34</td><td>Bag</td><td>@Zak_Bagans pets r like | r ?? @Zak_Bagans http:/ | @Zak_Bagans this is Sab | e arrived Bago | #Flood in Bago Myanmar | ts Zipper Bags Coffee h | Sho</td></tr>
<tr><td>6.94%</td><td>70</td><td>5</td><td>67</td><td>bag</td><td>ece of cabbage????????? | hiking garbage-bot (des | dio @Heavybag201 @battl | ‰Ã›Ã·body bagsÂ‰Ã›Âª - | Body Handbags for Wome | a Vickers bags: machi</td></tr>
</table>

Parameter Settings: Maximum Target Mean=.2, Maximum Target Mean=.8, Minimum Document Frequency=10

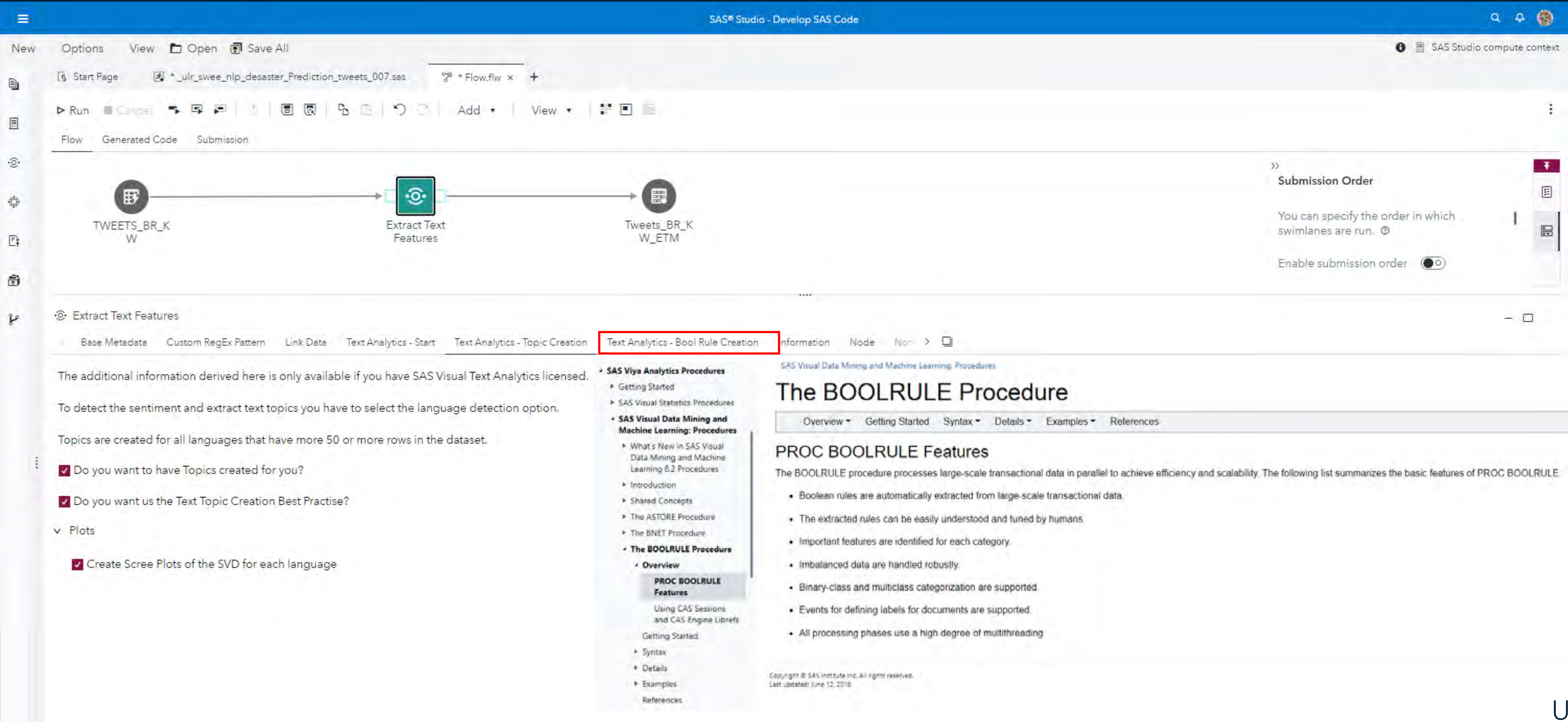


SAS® Studio - Develop SAS Code
New
Options
View
Open
Save All
SAS Studio compute context
Start Page
* _ulr_swee_nlp_desaster_Prediction_tweets_007.sas
* Flow.flw
Run
Cancel
Add
View
Flow
Generated Code
Submission
TWEETS_BR_KW
Extract Text Features
Tweets_BR_KW_ETM
Submission Order
You can specify the order in which swimlanes are run.
Extract Text Features
Base Metadata
Custom RegEx Pattern
Link Data
Text Analytics - Start
Text Analytics - Topic Creation
Text Analytics - Bool Rule C
Collect additional information from the links in the tweets. This option requires you to have enabled the Options for concatenated and separated columns.
Please note for this step to work your environment needs to be able to make calls to the open internet.
Please be also aware that this step can take a loot of time to run as the individual sites have to be called and their output need to be parsed.
The following five features are extracted for each link:
- HTTP Status Code (basically is it reachable or not)
- Title of the Webpage
- Description of the Webpage
- URL of the Webpage (handy if URL shorteners were used
- Owner of the site
Do you want to collect metadata from Links in the text?
Do you want to allow Unverified Requests (Warning potential impact: Breach of Confidentiality & Breach of Integrity)?
Flow.flw [temp]
Recover (3)
Submission (0)


SAS® Studio - Develop SAS Code
New
Options
View
Open
Save All
SAS Studio compute context
Start Page
*_ulr_swee_nlp_desaster_Prediction_tweets_007.sas
* Flow.flw
Run
Cancel
Add
View
Flow
Generated Code
Submission
TWEETS_BR_K W
Extract Text Features
Tweets_BR_K W_ETM
Submission Order
You can specify the order in which swimlanes are run.
Enable submission order
Extract Text Features
Base Metadata
Custom RegEx Pattern
Link Data
Text Analytics - Start
Text Analytics - Topic Creation
Text Analytics - Bool Rule Creation
Information
Node
The additional information derived here is only available if you have SAS Visual Text Analytics licensed.
To detect the sentiment and extract text topics you have to select the language detection option.
Do you want to use Text Analytics? (license required)
Do you want to automatically detect the text language?
Yes
No
Please select the language of your text:
English
Text Profiling
Do you want to profile your text?
Compare your text corpus to reference corpus profiles (Not available for all languages yet, raises a warning accordingly)
Add Word and Sentence count per Document and Language to the Results
Create a feature for the number of sentences in the Text
Create a feature for the count of tokens in the longest sentence
Sentiment detection
Do you want to detect the text sentiment?
Create Plot of Sentiment by Languages
Text Concept Extraction
Selecting SAS Predefined Concepts also enables you to use these as features in the customization options for the Text Topic Creation.
Do you want to apply the SAS Predefined Concepts?
Select Predefined Concepts you wish applie... 9 sele...
Noun Group
Organization
Percent
Person
Place
Time
Create Plot of Extracted SAS Predefined Concepts for each Language
Do you want to create a concated column of all matched text for each pre defined concept type?
Custom Text Concept Extraction
Please ensure that you have validated your Custom Concepts before using this step.
To add the table containing your Custom Concepts right click the step in the flow > Expand Input Ports and then connect your desired input table to the new
Supplying your own Custom Concepts also enables you to use these features in the customization options for the Text Topic Creation.
Do you want to apply custom concepts?
Flow.flw [temp]
Recover (3)
Submission (0

Derive Predictors from

Boolean Term Rules



SAS® Studio - Develop SAS Code

New Options View Open Save All SAS Studio compute context

Start Page | *_ulr_swee_nlp_desaster_Prediction_tweets_007.sas | * Flow.flw

Run Cancel Add View

Flow | Generated Code | Submission

TWEETS_BR_KW → Extract Text Features → Tweets_BR_KW_ETM

Submission Order

You can specify the order in which swimlanes are run.

Extract Text Features

Base Metadata | Custom RegEx Pattern | Link Data | Text Analytics - Start | Text Analytics - Topic Creation | Text Analytics - Bool Rule Creation | Information | Node | Notes

☑ Do you want to leverage BoolRule Creation to create additional Features?

☑ Do you want use the Term-by-Document Matrix from the Topic Creation?

Please select a target variable: *

target

Please select the Type of your Target Variable: *

Binary

Customization

Enter the Minimum G-Score needed for a Positive Term to be considered for Rule Extraction:

10

Enter the M Value for computing Estimated Precision for Positive Terms:

10

Enter the Minimum G-Score needed for a Negative Term to be considered for Rule Extraction:

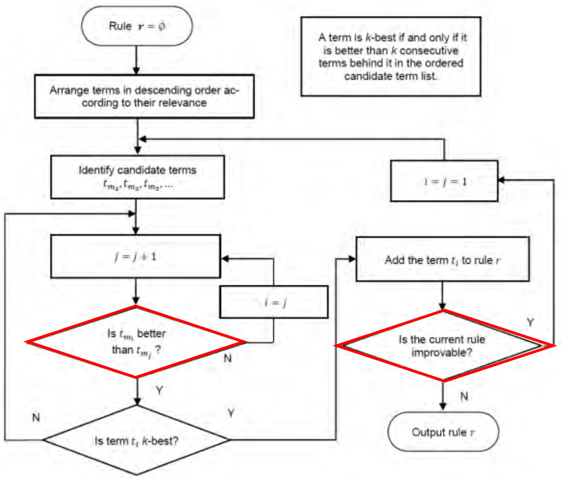
10

Enter the M Value for computing Estimated Precision for Negative Terms:

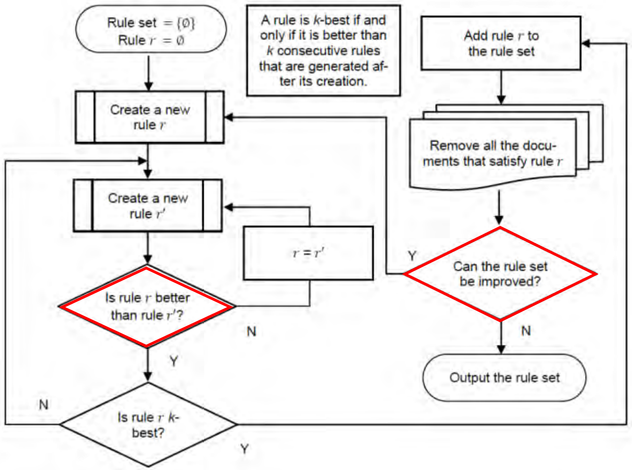
10

Enter the Minimum Number of Documents in which a Term ...

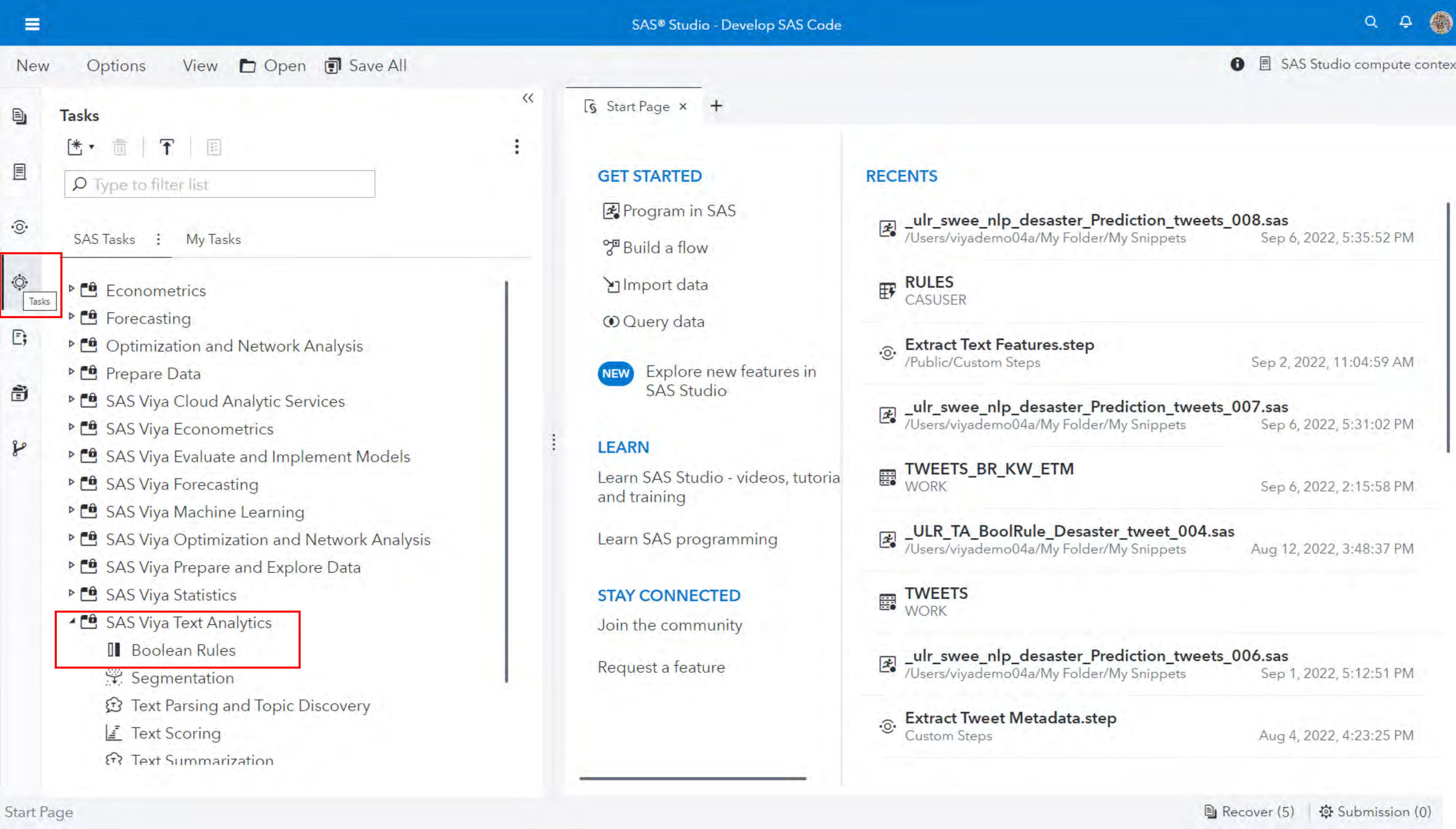
Term Ensemble Process for Creating a Rule

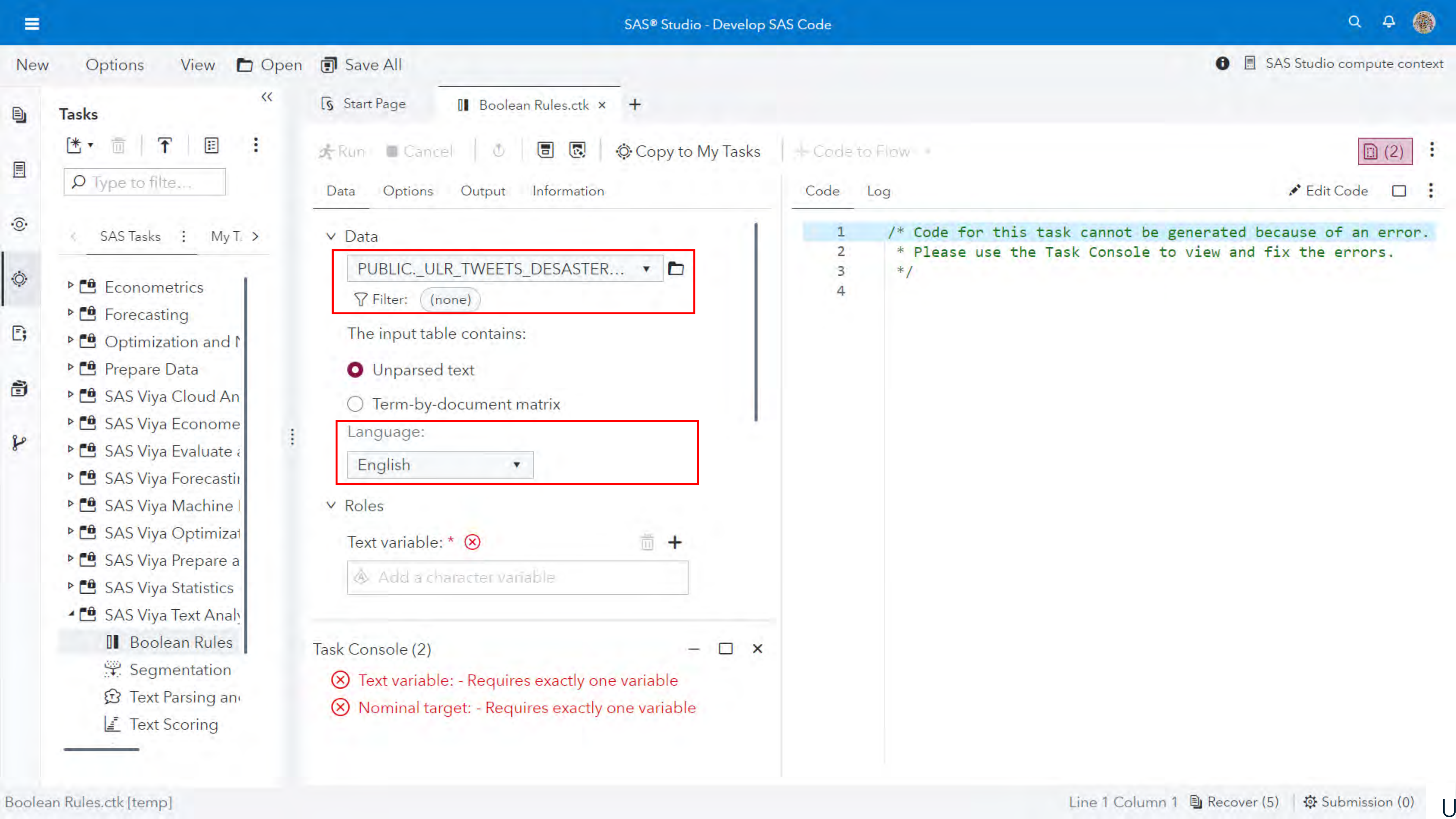


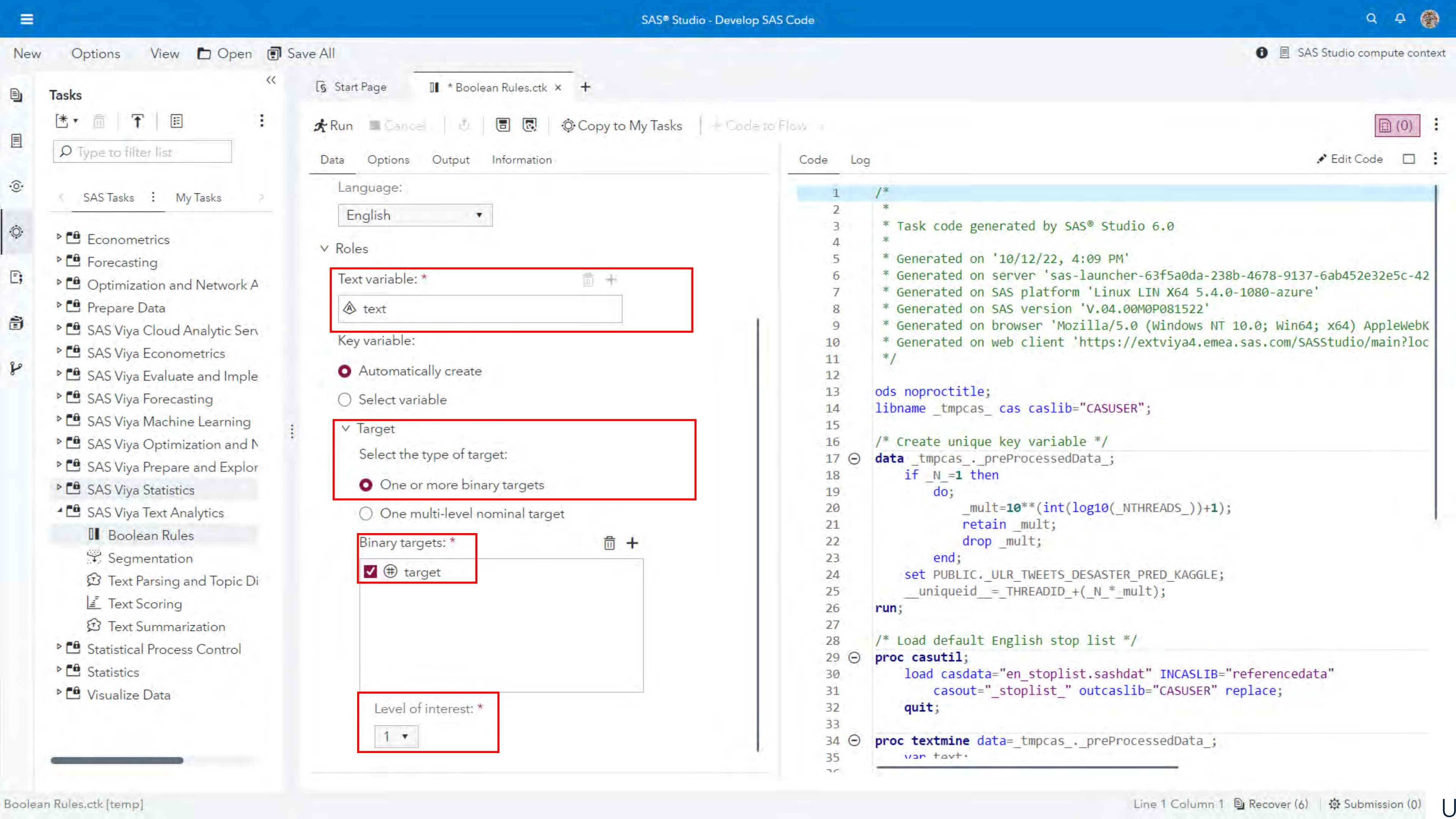
Rule Ensemble for Creating a Rule Set

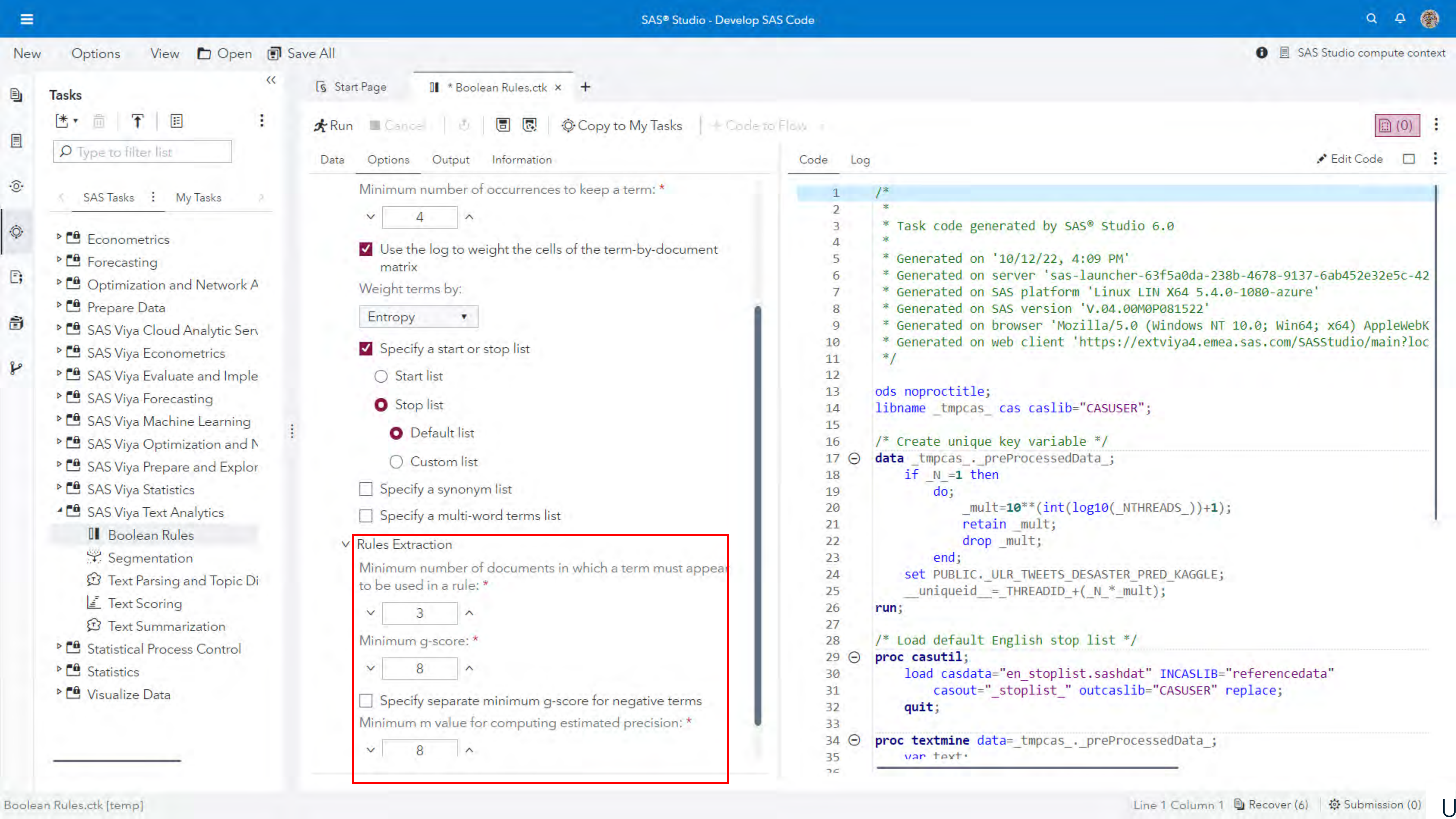


Flow.flw [temp] Recover (3) Submission (0)

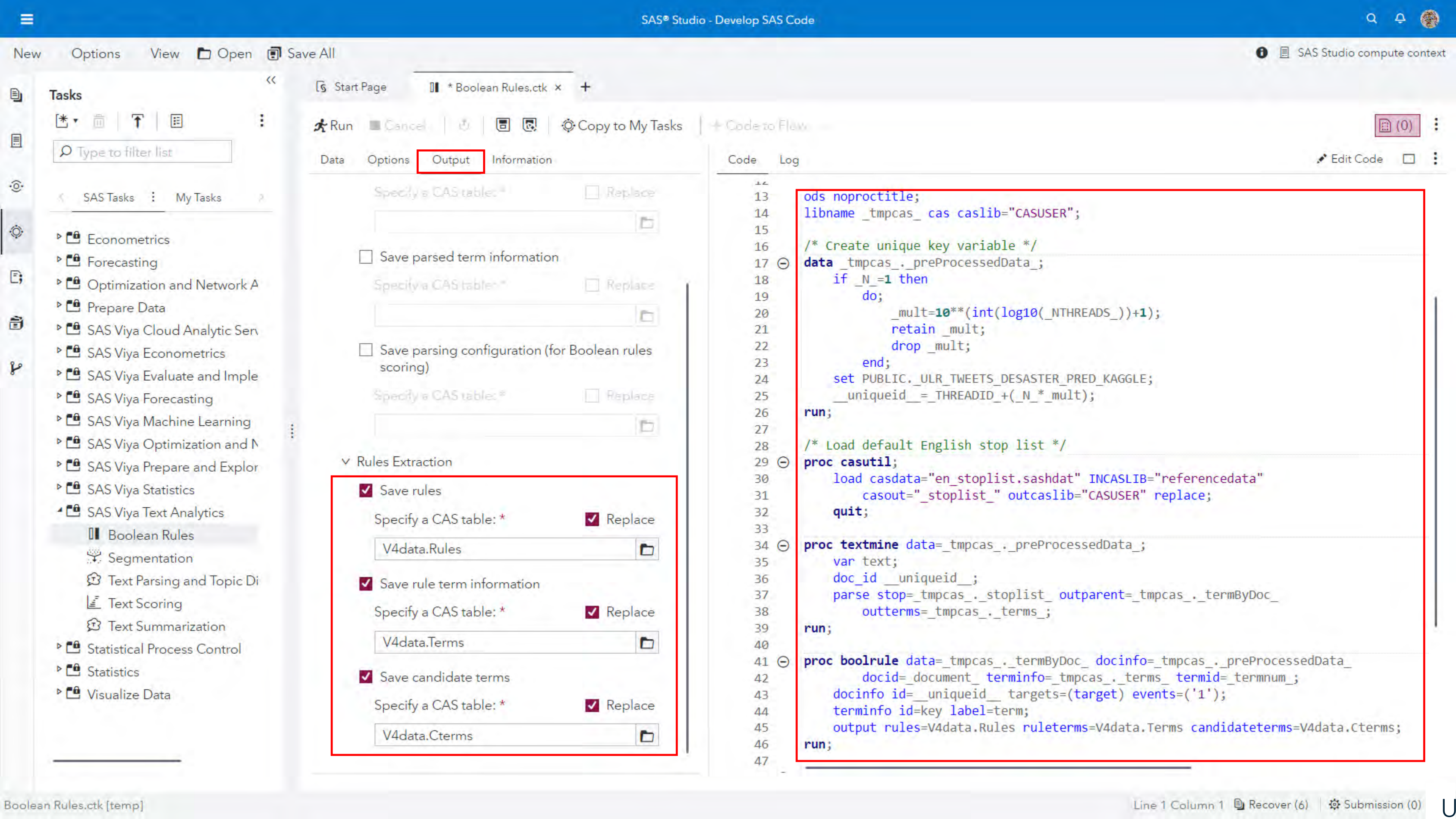

SAS® Studio - Develop SAS Code
New
Options
View
Open
Save All
SAS Studio compute context
Tasks
Type to filter list
SAS Tasks
My Tasks
Tasks
Econometrics
Forecasting
Optimization and Network Analysis
Prepare Data
SAS Viya Cloud Analytic Services
SAS Viya Econometrics
SAS Viya Evaluate and Implement Models
SAS Viya Forecasting
SAS Viya Machine Learning
SAS Viya Optimization and Network Analysis
SAS Viya Prepare and Explore Data
SAS Viya Statistics
SAS Viya Text Analytics
Boolean Rules
Segmentation
Text Parsing and Topic Discovery
Text Scoring
Text Summarization
Start Page
GET STARTED
Program in SAS
Build a flow
Import data
Query data
NEW
Explore new features in SAS Studio
LEARN
Learn SAS Studio - videos, tutoria and training
Learn SAS programming
STAY CONNECTED
Join the community
Request a feature
RECENTS
_ulr_swee_nlp_desaster_Prediction_tweets_008.sas
/Users/viyademo04a/My Folder/My Snippets
Sep 6, 2022, 5:35:52 PM
RULES
CASUSER
Extract Text Features.step
/Public/Custom Steps
Sep 2, 2022, 11:04:59 AM
_ulr_swee_nlp_desaster_Prediction_tweets_007.sas
/Users/viyademo04a/My Folder/My Snippets
Sep 6, 2022, 5:31:02 PM
TWEETS_BR_KW_ETM
WORK
Sep 6, 2022, 2:15:58 PM
_ULR_TA_BoolRule_Desaster_tweet_004.sas
/Users/viyademo04a/My Folder/My Snippets
Aug 12, 2022, 3:48:37 PM
TWEETS
WORK
_ulr_swee_nlp_desaster_Prediction_tweets_006.sas
/Users/viyademo04a/My Folder/My Snippets
Sep 1, 2022, 5:12:51 PM
Extract Tweet Metadata.step
Custom Steps
Aug 4, 2022, 4:23:25 PM
Start Page
Recover (5)
Submission (0)
U

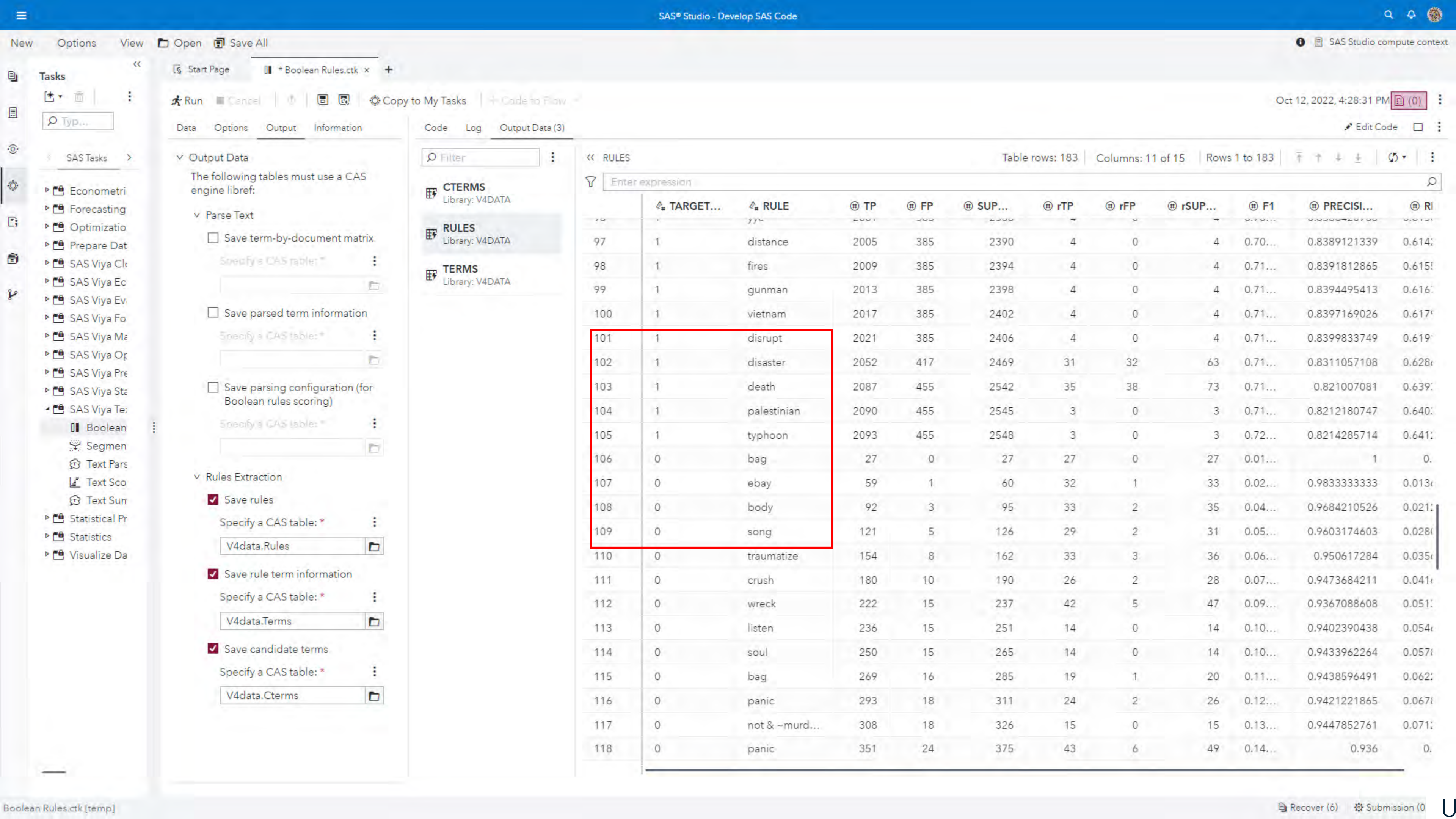

SAS® Studio - Develop SAS Code
New
Options
View
Open
Save All
SAS Studio compute context
Tasks
Type to filte...
SAS Tasks
My T.
Econometrics
Forecasting
Optimization and I
Prepare Data
SAS Viya Cloud An
SAS Viya Econome
SAS Viya Evaluate
SAS Viya Forecasti
SAS Viya Machine
SAS Viya Optimiza
SAS Viya Prepare a
SAS Viya Statistics
SAS Viya Text Anal
Boolean Rules
Segmentation
Text Parsing an
Text Scoring
Start Page
Boolean Rules.ctk
Run
Cancel
Copy to My Tasks
Code to Flow
(2)
Data
Options
Output
Information
Data
PUBLIC._ULR_TWEETS_DESASTER...
Filter:
(none)
The input table contains:
Unparsed text
Term-by-document matrix
Language:
English
Roles
Text variable: *
Add a character variable
Task Console (2)
Text variable: - Requires exactly one variable
Nominal target: - Requires exactly one variable
Code
Log
Edit Code
1 /* Code for this task cannot be generated because of an error.
2 * Please use the Task Console to view and fix the errors.
3 */
4
Boolean Rules.ctk [temp]
Line 1 Column 1
Recover (5)
Submission (0)


SAS® Studio - Develop SAS Code
New
Options
View
Open
Save All
SAS Studio compute context
Tasks
Type to filter list
SAS Tasks
My Tasks
Econometrics
Forecasting
Optimization and Network A
Prepare Data
SAS Viya Cloud Analytic Serv
SAS Viya Econometrics
SAS Viya Evaluate and Imple
SAS Viya Forecasting
SAS Viya Machine Learning
SAS Viya Optimization and N
SAS Viya Prepare and Explor
SAS Viya Statistics
SAS Viya Text Analytics
Boolean Rules
Segmentation
Text Parsing and Topic Di
Text Scoring
Text Summarization
Statistical Process Control
Statistics
Visualize Data
Start Page
* Boolean Rules.ctk
Run
Cancel
Copy to My Tasks
Code to Flow
(0)
Data
Options
Output
Information
Language:
English
Roles
Text variable: *
text
Key variable:
Automatically create
Select variable
Target
Select the type of target:
One or more binary targets
One multi-level nominal target
Binary targets: *
target
Level of interest: *
1
Code
Log
Edit Code
/*
 *
 * Task code generated by SAS® Studio 6.0
 *
 * Generated on '10/12/22, 4:09 PM'
 * Generated on server 'sas-launcher-63f5a0da-238b-4678-9137-6ab452e32e5c-42
 * Generated on SAS platform 'Linux LIN X64 5.4.0-1080-azure'
 * Generated on SAS version 'V.04.00M0P081522'
 * Generated on browser 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebK
 * Generated on web client 'https://extviya4.emea.sas.com/SASStudio/main?loc
 */
ods noproctitle;
libname _tmpcas_ cas caslib="CASUSER";
/* Create unique key variable */
data _tmpcas_._preProcessedData_;
if _N_=1 then
do;
_mult=10**(int(log10(_NTHREADS_))+1);
retain _mult;
drop _mult;
end;
set PUBLIC._ULR_TWEETS_DESASTER_PRED_KAGGLE;
__uniqueid__=_THREADID_+(_N_*_mult);
run;
/* Load default English stop list */
proc casutil;
load casdata="en_stoplist.sashdat" INCASLIB="referencedata"
casout="_stoplist_" outcaslib="CASUSER" replace;
quit;
proc textmine data=_tmpcas_._preProcessedData_;
Boolean Rules.ctk [temp]
Line 1 Column 1
Recover (6)
Submission (0)

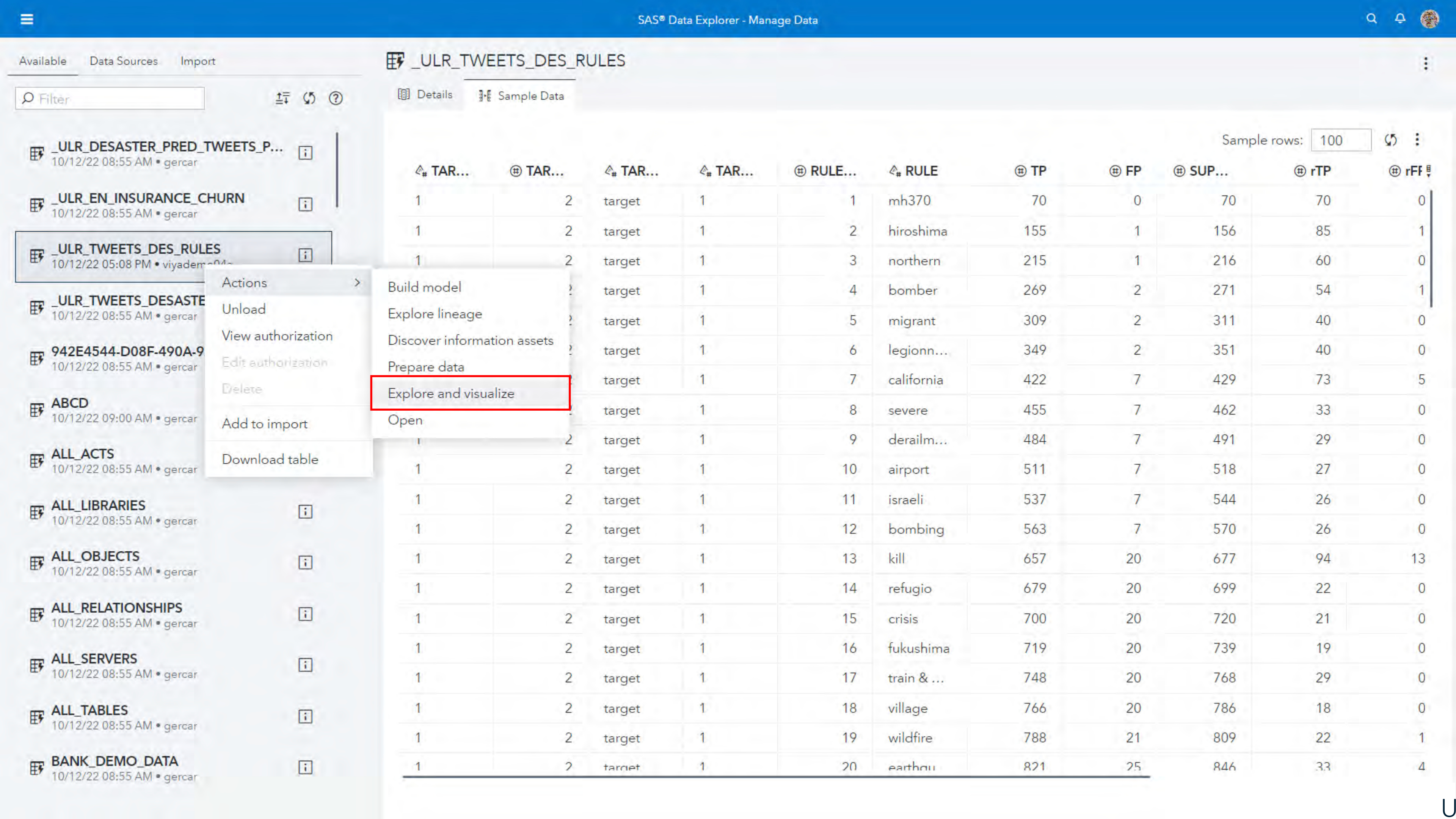


SAS® Studio - Develop SAS Code
New
Options
View
Open
Save All
SAS Studio compute context
Tasks
Type to filter list
SAS Tasks
My Tasks
Econometrics
Forecasting
Optimization and Network A
Prepare Data
SAS Viya Cloud Analytic Serv
SAS Viya Econometrics
SAS Viya Evaluate and Imple
SAS Viya Forecasting
SAS Viya Machine Learning
SAS Viya Optimization and N
SAS Viya Prepare and Explor
SAS Viya Statistics
SAS Viya Text Analytics
Boolean Rules
Segmentation
Text Parsing and Topic Di
Text Scoring
Text Summarization
Statistical Process Control
Statistics
Visualize Data
Start Page
* Boolean Rules.ctk
Run
Cancel
Copy to My Tasks
Code to Flow
(0)
Data
Options
Output
Information
Minimum number of occurrences to keep a term: *
4
Use the log to weight the cells of the term-by-document matrix
Weight terms by:
Entropy
Specify a start or stop list
Start list
Stop list
Default list
Custom list
Specify a synonym list
Specify a multi-word terms list
Rules Extraction
Minimum number of documents in which a term must appear to be used in a rule: *
3
Minimum g-score: *
8
Specify separate minimum g-score for negative terms
Minimum m value for computing estimated precision: *
8
Code
Log
Edit Code
/*
*
* Task code generated by SAS® Studio 6.0
*
* Generated on '10/12/22, 4:09 PM'
* Generated on server 'sas-launcher-63f5a0da-238b-4678-9137-6ab452e32e5c-42
* Generated on SAS platform 'Linux LIN X64 5.4.0-1080-azure'
* Generated on SAS version 'V.04.00M0P081522'
* Generated on browser 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebK
* Generated on web client 'https://extviya4.emea.sas.com/SASStudio/main?loc
*/
ods noproctitle;
libname _tmpcas_ cas caslib="CASUSER";
/* Create unique key variable */
data _tmpcas_._preProcessedData_;
if _N_=1 then
do;
_mult=10**(int(log10(_NTHREADS_))+1);
retain _mult;
drop _mult;
end;
set PUBLIC._ULR_TWEETS_DESASTER_PRED_KAGGLE;
__uniqueid__=_THREADID_+(_N_*_mult);
run;
/* Load default English stop list */
proc casutil;
load casdata="en_stoplist.sashdat" INCASLIB="referencedata"
casout="_stoplist_" outcaslib="CASUSER" replace;
quit;
proc textmine data=_tmpcas_._preProcessedData_;
Boolean Rules.ctk [temp]
Line 1 Column 1
Recover (6)
Submission (0)


SAS® Studio - Develop SAS Code
New
Options
View
Open
Save All
SAS Studio compute context
Tasks
Type to filter list
SAS Tasks
My Tasks
Econometrics
Forecasting
Optimization and Network A
Prepare Data
SAS Viya Cloud Analytic Serv
SAS Viya Econometrics
SAS Viya Evaluate and Imple
SAS Viya Forecasting
SAS Viya Machine Learning
SAS Viya Optimization and N
SAS Viya Prepare and Explor
SAS Viya Statistics
SAS Viya Text Analytics
Boolean Rules
Segmentation
Text Parsing and Topic Di
Text Scoring
Text Summarization
Statistical Process Control
Statistics
Visualize Data
Start Page
* Boolean Rules.ctk
Run
Cancel
Copy to My Tasks
Code to Flow
(0)
Data
Options
Output
Information
Specify a CAS table: *
Replace
Save parsed term information
Specify a CAS table: *
Replace
Save parsing configuration (for Boolean rules scoring)
Specify a CAS table: *
Replace
Rules Extraction
Save rules
Specify a CAS table: *
Replace
V4data.Rules
Save rule term information
Specify a CAS table: *
Replace
V4data.Terms
Save candidate terms
Specify a CAS table: *
Replace
V4data.Cterms
Code
Log
Edit Code
ods noproctitle;
libname _tmpcas_ cas caslib="CASUSER";
/* Create unique key variable */
data _tmpcas_._preProcessedData_;
if _N_=1 then
do;
_mult=10**(int(log10(_NTHREADS_))+1);
retain _mult;
drop _mult;
end;
set PUBLIC._ULR_TWEETS_DESASTER_PRED_KAGGLE;
__uniqueid__=_THREADID_+(_N_*_mult);
run;
/* Load default English stop list */
proc casutil;
load casdata="en_stoplist.sashdat" INCASLIB="referencedata"
casout="_stoplist_" outcaslib="CASUSER" replace;
quit;
proc textmine data=_tmpcas_._preProcessedData_;
var text;
doc_id __uniqueid__;
parse stop=_tmpcas_._stoplist_ outparent=_tmpcas_._termByDoc_
outterms=_tmpcas_._terms_;
run;
proc boolrule data=_tmpcas_._termByDoc_ docinfo=_tmpcas_._preProcessedData_
docid=_document_ terminfo=_tmpcas_._terms_ termid=_termnum_;
docinfo id=__uniqueid__ targets=(target) events=('1');
terminfo id=key label=term;
output rules=V4data.Rules ruleterms=V4data.Terms candidateterms=V4data.Cterms;
run;
Boolean Rules.ctk [temp]
Line 1 Column 1
Recover (6)
Submission (0)



	TARGET...	RULE	TP	FP	SUP...	rTP	rFP	rSUP...	F1	PRECISI...	RI
97	1	distance	2005	385	2390	4	0	4	0.70...	0.8389121339	0.614
98	1	fires	2009	385	2394	4	0	4	0.71...	0.8391812865	0.615
99	1	gunman	2013	385	2398	4	0	4	0.71...	0.8394495413	0.616
100	1	vietnam	2017	385	2402	4	0	4	0.71...	0.8397169026	0.617
101	1	disrupt	2021	385	2406	4	0	4	0.71...	0.8399833749	0.619
102	1	disaster	2052	417	2469	31	32	63	0.71...	0.8311057108	0.628
103	1	death	2087	455	2542	35	38	73	0.71...	0.821007081	0.639
104	1	palestinian	2090	455	2545	3	0	3	0.71...	0.8212180747	0.640
105	1	typhoon	2093	455	2548	3	0	3	0.72...	0.8214285714	0.641
106	0	bag	27	0	27	27	0	27	0.01...	1	0.
107	0	ebay	59	1	60	32	1	33	0.02...	0.9833333333	0.013
108	0	body	92	3	95	33	2	35	0.04...	0.9684210526	0.021
109	0	song	121	5	126	29	2	31	0.05...	0.9603174603	0.028
110	0	traumatize	154	8	162	33	3	36	0.06...	0.950617284	0.035
111	0	crush	180	10	190	26	2	28	0.07...	0.9473684211	0.041
112	0	wreck	222	15	237	42	5	47	0.09...	0.9367088608	0.051
113	0	listen	236	15	251	14	0	14	0.10...	0.9402390438	0.054
114	0	soul	250	15	265	14	0	14	0.10...	0.9433962264	0.057
115	0	bag	269	16	285	19	1	20	0.11...	0.9438596491	0.062
116	0	panic	293	18	311	24	2	26	0.12...	0.9421221865	0.067
117	0	not & ~murd...	308	18	326	15	0	15	0.13...	0.9447852761	0.071
118	0	panic	351	24	375	43	6	49	0.14...	0.936	0.


SAS® Data Explorer - Manage Data
Available
Data Sources
Import
Filter
_ULR_DESASTER_PRED_TWEETS_P...
10/12/22 08:55 AM • gercar
_ULR_EN_INSURANCE_CHURN
10/12/22 08:55 AM • gercar
_ULR_TWEETS_DES_RULES
10/12/22 05:08 PM • viyadem
_ULR_TWEETS_DESASTE
10/12/22 08:55 AM • gercar
942E4544-D08F-490A-9
10/12/22 08:55 AM • gercar
ABCD
10/12/22 09:00 AM • gercar
ALL_ACTS
10/12/22 08:55 AM • gercar
ALL_LIBRARIES
10/12/22 08:55 AM • gercar
ALL_OBJECTS
10/12/22 08:55 AM • gercar
ALL_RELATIONSHIPS
10/12/22 08:55 AM • gercar
ALL_SERVERS
10/12/22 08:55 AM • gercar
ALL_TABLES
10/12/22 08:55 AM • gercar
BANK_DEMO_DATA
10/12/22 08:55 AM • gercar
Actions
Unload
View authorization
Edit authorization
Delete
Add to import
Download table
Build model
Explore lineage
Discover information assets
Prepare data
Explore and visualize
Open
_ULR_TWEETS_DES_RULES
Details
Sample Data
Sample rows: 100
TAR... TAR... TAR... TAR... RULE... RULE TP FP SUP... rTP rFP
1 2 target 1 1 mh370 70 0 70 70 0
1 2 target 1 2 hiroshima 155 1 156 85 1
target 1 3 northern 215 1 216 60 0
target 1 4 bomber 269 2 271 54 1
target 1 5 migrant 309 2 311 40 0
target 1 6 legionn... 349 2 351 40 0
target 1 7 california 422 7 429 73 5
target 1 8 severe 455 7 462 33 0
target 1 9 derailm... 484 7 491 29 0
1 2 target 1 10 airport 511 7 518 27 0
1 2 target 1 11 israeli 537 7 544 26 0
1 2 target 1 12 bombing 563 7 570 26 0
1 2 target 1 13 kill 657 20 677 94 13
1 2 target 1 14 refugio 679 20 699 22 0
1 2 target 1 15 crisis 700 20 720 21 0
1 2 target 1 16 fukushima 719 20 739 19 0
1 2 target 1 17 train & ... 748 20 768 29 0
1 2 target 1 18 village 766 20 786 18 0
1 2 target 1 19 wildfire 788 21 809 22 1
1 2 target 1 20 earthqu 821 25 846 33 4

SAS® Visual Analytics - Explore and Visualize

Editing

Bolean Rule Result Browser (NLP Desaster Prediction Twitter)

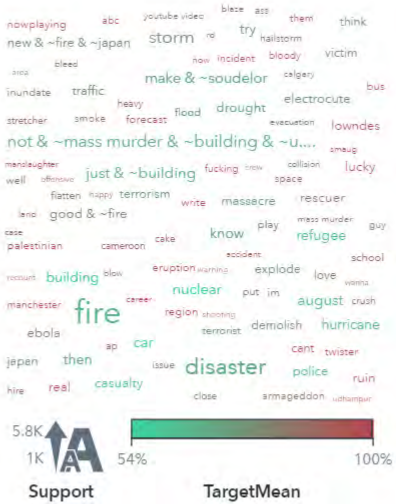
Page 1

Drag a data item or control here to create a page prompt.

Filters: 81.5%; 100.0% > 6; 318 > 1.2%; 71.9% > 0.6%; 63.9%

Frequency	Total Precision	Target Value	Cumulative Precision	Cumulative F1 - Score	Cumulativs Recall	Min Rule Support
162	0.82	0 / 1	81.5% to 100.0% (81.5% – 100.0%)	1.2% to 71.9% (1.2% – 71.9%)	0.6% to 63.9% (0.6% – 63.9%)	6 to 318 (3 – 31)

Support, TargetMean by rule



rID ▲	target	rule	rSupport	TargetMean	rTP	rFP	FP	TP	precision	F-1 score	recall	Support
1	1	mh370	70	100.0%	70	0	0	70	100.0%	4.2%	2.1%	70
2	1	hiroshima	86	98.8%	85	1	1	155	99.4%	9.1%	4.7%	156
3	1	northern	60	100.0%	60	0	1	215	99.5%	12.4%	6.6%	216
4	1	bomber	55	98.2%	54	1	2	269	99.3%	15.2%	8.2%	271
5	1	migrant	40	100.0%	40	0	2	309	99.4%	17.3%	9.5%	311
6	1	legionnaire	40	100.0%	40	0	2	349	99.4%	19.3%	10.7%	351
7	1	california	78	93.6%	73	5	7	422	98.4%	22.9%	12.9%	429
8	1	severe	33	100.0%	33	0	7	455	98.5%	24.4%	13.9%	462
9	1	derailment	29	100.0%	29	0	7	484	98.6%	25.8%	14.8%	491
10	1	airport	27	100.0%	27	0	7	511	98.6%	27.0%	15.7%	518
11	1	israeli	26	100.0%	26	0	7	537	98.7%	28.2%	16.5%	544
12	1	bombing	26	100.0%	26	0	7	563	98.8%	29.4%	17.2%	570
13	1	kill	107	87.9%	94	13	20	657	97.0%	33.3%	20.1%	677
14	1	refugio	22	100.0%	22	0	20	679	97.1%	34.3%	20.8%	699
15	1	crisis	21	100.0%	21	0	20	700	97.2%	35.1%	21.4%	720
16	1	fukushima	19	100.0%	19	0	20	719	97.3%	35.9%	22.0%	739
17	1	train & derail	29	100.0%	29	0	20	748	97.4%	37.1%	22.9%	768
18	1	village	18	100.0%	18	0	20	766	97.5%	37.8%	23.5%	786

SAS® Visual Analytics - Explore and Visualize

Editing

Bolean Rule Result Browser (NLP Desaster Prediction Twitter)

Page 1

Drag a data item or control here to create a page prompt.

Filters: 81.5%; 100.0% > 6; 318 > 1.2%; 71.9% > 0.6%; 63.9% > 1 ×

Frequency 84

Total Precision 0.83

Target Value: 0 | 1

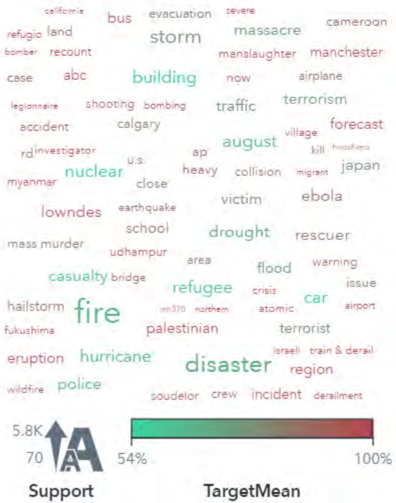
Cumulative Precision: 81.5% to 100.0% (81.5% – 100.0%)

Cumulative F1 - Score: 1.2% to 71.9% (1.2% – 71.9%)

Cumulativs Recall: 0.6% to 63.9% (0.6% – 63.9%)

Min Rule Support: 6 to 318 (3 – 31)

Support, TargetMean by rule



rID ▲	target	rule	rSupport	TargetMean	rTP	rFP	FP	TP	precision	F-1 score	recall	Suppor
1	1	mh370	70	100.0%	70	0	0	70	100.0%	4.2%	2.1%	70
2	1	hiroshima	86	98.8%	85	1	1	155	99.4%	9.1%	4.7%	156
3	1	northern	60	100.0%	60	0	1	215	99.5%	12.4%	6.6%	216
4	1	bomber	55	98.2%	54	1	2	269	99.3%	15.2%	8.2%	271
5	1	migrant	40	100.0%	40	0	2	309	99.4%	17.3%	9.5%	311
6	1	legionnaire	40	100.0%	40	0	2	349	99.4%	19.3%	10.7%	351
7	1	california	78	93.6%	73	5	7	422	98.4%	22.9%	12.9%	429
8	1	severe	33	100.0%	33	0	7	455	98.5%	24.4%	13.9%	462
9	1	derailment	29	100.0%	29	0	7	484	98.6%	25.8%	14.8%	491
10	1	airport	27	100.0%	27	0	7	511	98.6%	27.0%	15.7%	518
11	1	israeli	26	100.0%	26	0	7	537	98.7%	28.2%	16.5%	544
12	1	bombing	26	100.0%	26	0	7	563	98.8%	29.4%	17.2%	570
13	1	kill	107	87.9%	94	13	20	657	97.0%	33.3%	20.1%	677
14	1	refugio	22	100.0%	22	0	20	679	97.1%	34.3%	20.8%	699
15	1	crisis	21	100.0%	21	0	20	700	97.2%	35.1%	21.4%	720
16	1	fukushima	19	100.0%	19	0	20	719	97.3%	35.9%	22.0%	739
17	1	train & derail	29	100.0%	29	0	20	748	97.4%	37.1%	22.9%	768
18	1	village	18	100.0%	18	0	20	766	97.5%	37.8%	23.5%	786

SAS® Visual Analytics - Explore and Visualize

Editing

Bolean Rule Result Browser (NLP Desaster Prediction Twitter)

Page 1

Data | Objects | Suggest | Outline | Review

Options | Roles | Actions | Rules | Filters | Ranks

Drag a data item or control here to create a page prompt.

Filters: 81.5%; 100.0% > 6; 318 > 1.2%; 71.9% > 0.6%; 63.9% > 0 ×

Frequency 78

Total Precision 0.82

Target Value: 0 | 1

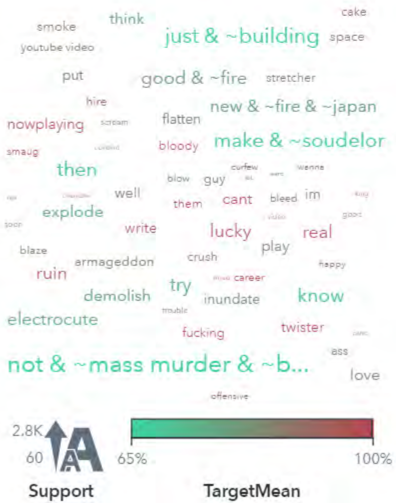
Cumulative Precision: 81.5% to 100.0% (81.5% – 100.0%)

Cumulative F1 - Score: 1.2% to 71.9% (1.2% – 71.9%)

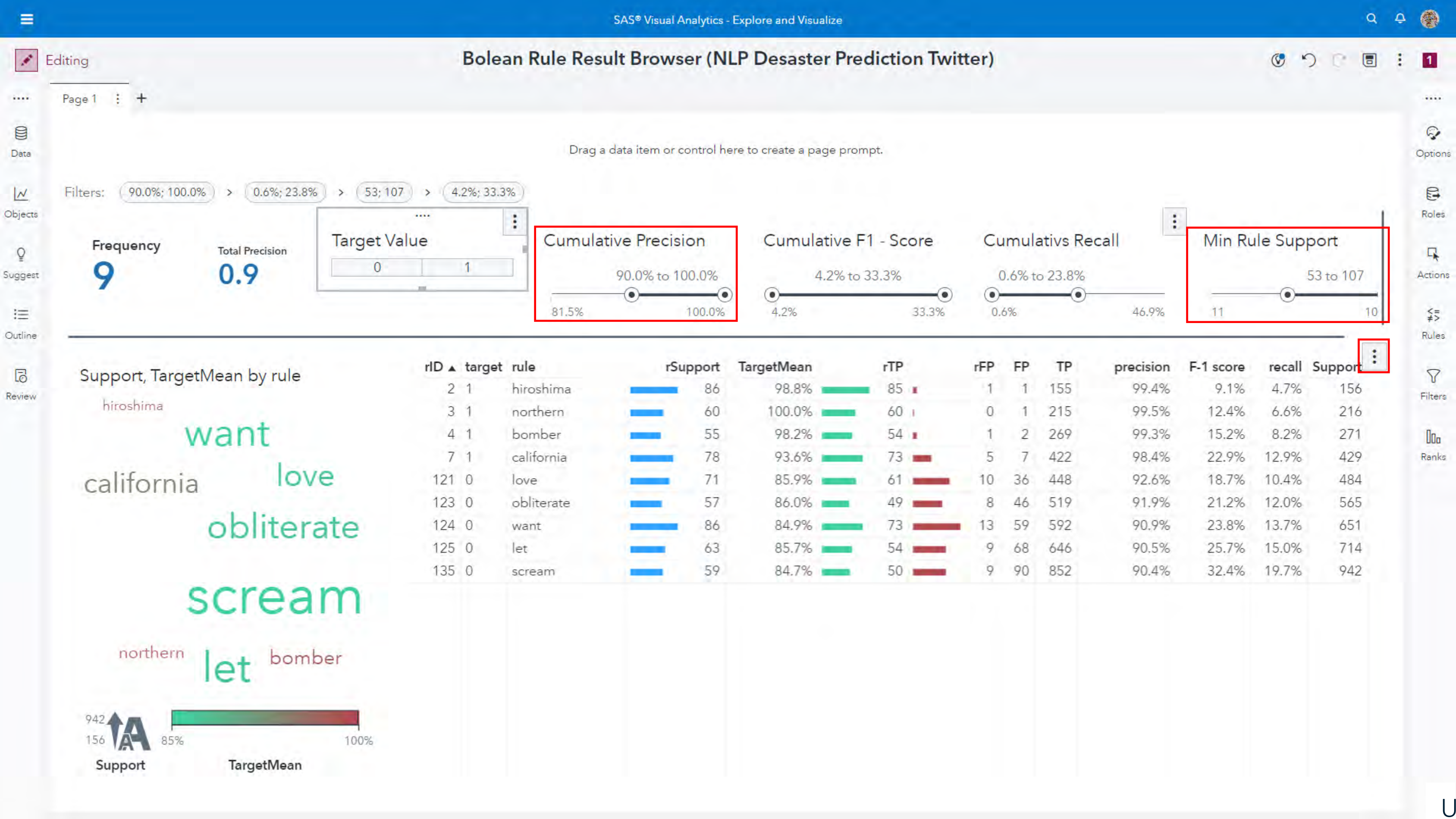
Cumulativs Recall: 0.6% to 63.9% (0.6% – 63.9%)

Min Rule Support: 6 to 318 (3 – 31)

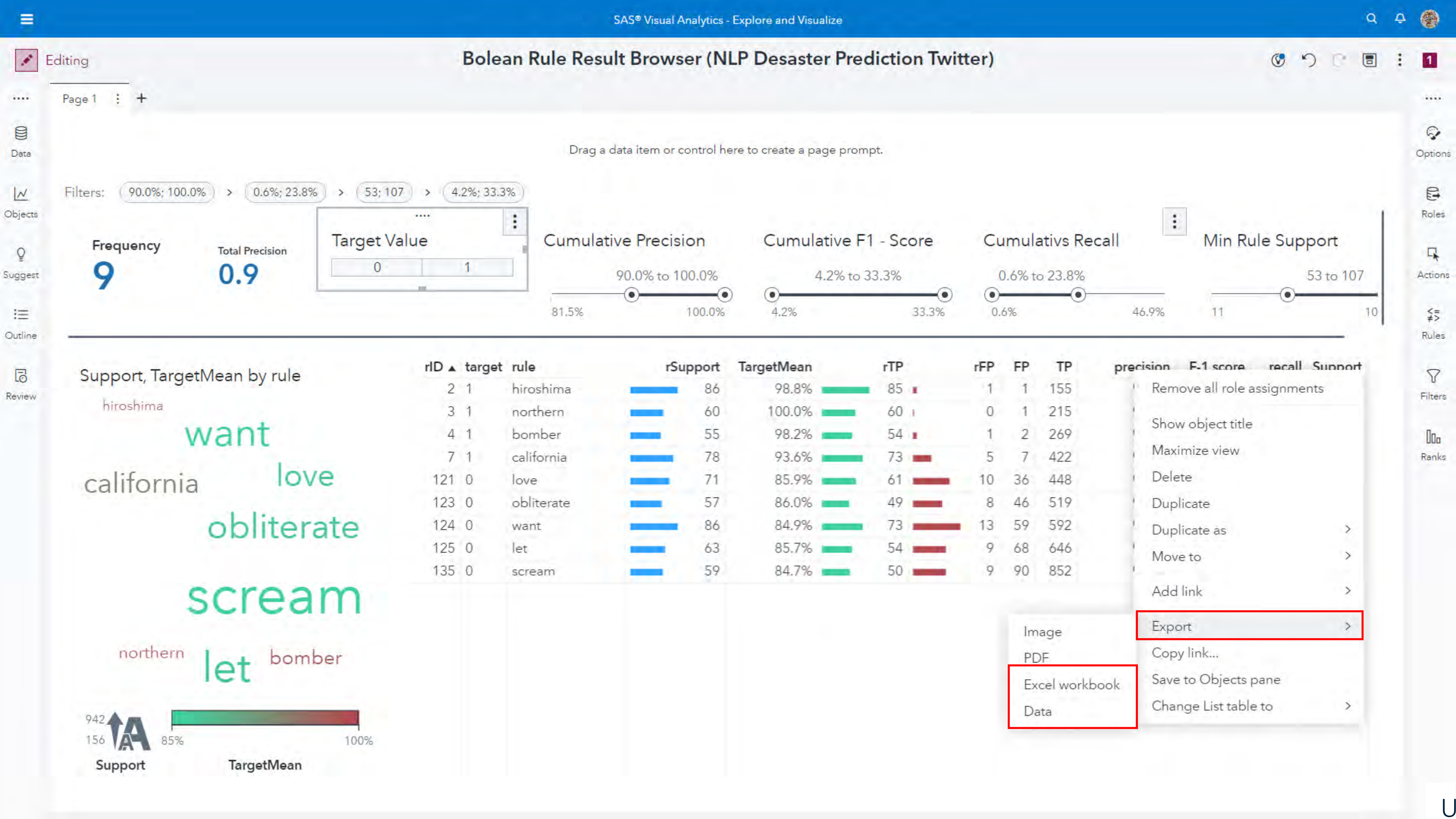
Support, TargetMean by rule



rID ▲	target	rule	rSupport	TargetMean	rTP	rFP	FP	TP	precision	F-1 score	recall	Support
107	0	ebay	33	97.0%	32	1	1	59	98.3%	2.7%	1.4%	60
108	0	body	35	94.3%	33	2	3	92	96.8%	4.2%	2.1%	95
109	0	song	31	93.5%	29	2	5	121	96.0%	5.4%	2.8%	126
110	0	traumatize	36	91.7%	33	3	8	154	95.1%	6.9%	3.6%	162
112	0	wreck	47	89.4%	42	5	15	222	93.7%	9.7%	5.1%	237
113	0	listen	14	100.0%	14	0	15	236	94.0%	10.3%	5.5%	251
114	0	soul	14	100.0%	14	0	15	250	94.3%	10.9%	5.8%	265
117	0	not & ~murder ...	15	100.0%	15	0	18	308	94.5%	13.3%	7.1%	326
119	0	aftershock	19	94.7%	18	1	25	369	93.7%	15.7%	8.5%	394
120	0	pick	19	94.7%	18	1	26	387	93.7%	16.4%	9.0%	413
122	0	obliteration	24	91.7%	22	2	38	470	92.5%	19.5%	10.9%	508
123	0	obliterate	57	86.0%	49	8	46	519	91.9%	21.2%	12.0%	565
124	0	want	86	84.9%	73	13	59	592	90.9%	23.8%	13.7%	651
125	0	let	63	85.7%	54	9	68	646	90.5%	25.7%	15.0%	714
126	0	reâ	23	91.3%	21	2	70	667	90.5%	26.4%	15.4%	737
127	0	character	12	100.0%	12	0	70	679	90.7%	26.8%	15.7%	749
128	0	king	12	100.0%	12	0	70	691	90.8%	27.2%	16.0%	761
129	0	content	22	90.9%	20	2	72	711	90.8%	27.9%	16.5%	783



rID ▲	target	rule	rSupport	TargetMean	rTP	rFP	FP	TP	precision	F-1 score	recall	Support
2	1	hiroshima	86	98.8%	85	1	1	155	99.4%	9.1%	4.7%	156
3	1	northern	60	100.0%	60	0	1	215	99.5%	12.4%	6.6%	216
4	1	bomber	55	98.2%	54	1	2	269	99.3%	15.2%	8.2%	271
7	1	california	78	93.6%	73	5	7	422	98.4%	22.9%	12.9%	429
121	0	love	71	85.9%	61	10	36	448	92.6%	18.7%	10.4%	484
123	0	obliterate	57	86.0%	49	8	46	519	91.9%	21.2%	12.0%	565
124	0	want	86	84.9%	73	13	59	592	90.9%	23.8%	13.7%	651
125	0	let	63	85.7%	54	9	68	646	90.5%	25.7%	15.0%	714
135	0	scream	59	84.7%	50	9	90	852	90.4%	32.4%	19.7%	942


SAS® Visual Analytics - Explore and Visualize
Editing
Bolean Rule Result Browser (NLP Desaster Prediction Twitter)
Page 1
Data
Objects
Suggest
Outline
Review
Options
Roles
Actions
Rules
Filters
Ranks
Drag a data item or control here to create a page prompt.
Filters: 90.0%; 100.0% > 0.6%; 23.8% > 53; 107 > 4.2%; 33.3%
Frequency
9
Total Precision
0.9
Target Value
0
1
Cumulative Precision
90.0% to 100.0%
81.5%
100.0%
Cumulative F1 - Score
4.2% to 33.3%
4.2%
33.3%
Cumulativs Recall
0.6% to 23.8%
0.6%
46.9%
Min Rule Support
53 to 107
11
10
Support, TargetMean by rule
hiroshima
want
california
love
obliterate
scream
northern
let
bomber
942
156
85%
100%
Support
TargetMean
rID target rule rSupport TargetMean rTP rFP FP TP precision F-1 score recall Support
2 1 hiroshima 86 98.8% 85 1 1 155
3 1 northern 60 100.0% 60 0 1 215
4 1 bomber 55 98.2% 54 1 2 269
7 1 california 78 93.6% 73 5 7 422
121 0 love 71 85.9% 61 10 36 448
123 0 obliterate 57 86.0% 49 8 46 519
124 0 want 86 84.9% 73 13 59 592
125 0 let 63 85.7% 54 9 68 646
135 0 scream 59 84.7% 50 9 90 852
Remove all role assignments
Show object title
Maximize view
Delete
Duplicate
Duplicate as
Move to
Add link
Export
Copy link...
Save to Objects pane
Change List table to
Image
PDF
Excel workbook
Data

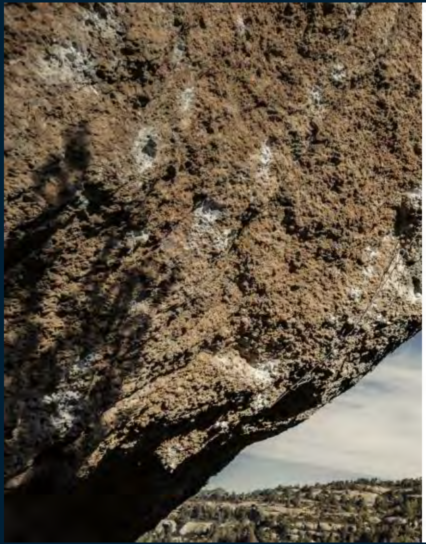
90.0% ≤ precision ≤ 100.0%
0.6% ≤ recall ≤ 23.8%
53 ≤ rSupport ≤ 107
4.2% ≤ F-1 score ≤ 33.3%

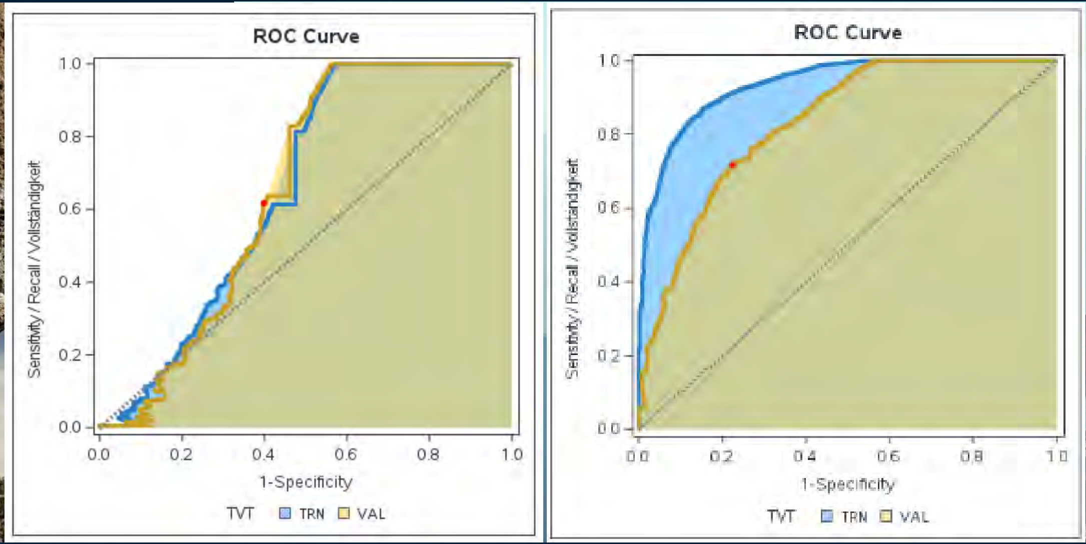
rID	target	rule	rSupport	TargetMean	rTP	rFP	FP	TP	precision	F-1 score	recall	Suppo
2	1	hiroshima	86	98,8%	85	1	1	155	99,4%	9,1%	4,7%	15
3	1	northern	60	100,0%	60	0	1	215	99,5%	12,4%	6,6%	21
4	1	bomber	55	98,2%	54	1	2	269	99,3%	15,2%	8,2%	27
7	1	california	78	93,6%	73	5	7	422	98,4%	22,9%	12,9%	42
121	0	love	71	85,9%	61	10	36	448	92,6%	18,7%	10,4%	48
123	0	obliterate	57	86,0%	49	8	46	519	91,9%	21,2%	12,0%	56
124	0	want	86	84,9%	73	13	59	592	90,9%	23,8%	13,7%	65
125	0	let	63	85,7%	54	9	68	646	90,5%	25,7%	15,0%	71
135	0	scream	59	84,7%	50	9	90	852	90,4%	32,4%	19,7%	94

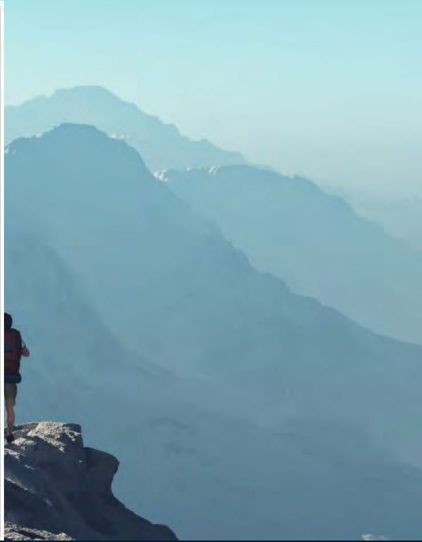
```
proc sql noprint;
    select ruleid , compress(translate(strip(rule),'|','&'),' ~')
           into : ranks separated by '|',
                : terms separated by '|'
    from tmpcas.rules
    order by ruleid;
Quit;

%let N=&sqlobs.;
%put &=N.;
%put &=ranks.;
%put &=terms.;
data ABT_BR;
    set ABT;
    %do i=1 %to &n.;
        BR_&i.=(find(&text.,"%scan(&terms.,&i.,'|')",'i') gt 0);
        Label BR_&i.="BR_&I.: %scan(&terms.,&i.,'|')";
    %end;
run;
```

This SQL and Data Step attaches binary variables for the selected terms indicating "1" whenever a term appears in the text.

Welche features verbessern hier die Klassifikation?







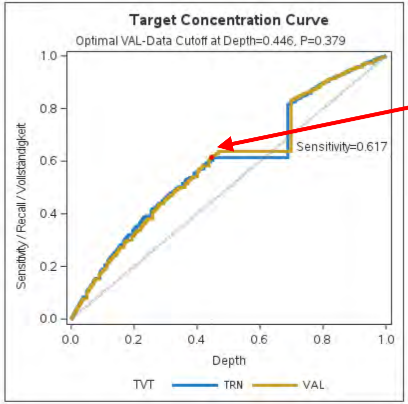
To answer this question, we need to attempt to climb the ROC

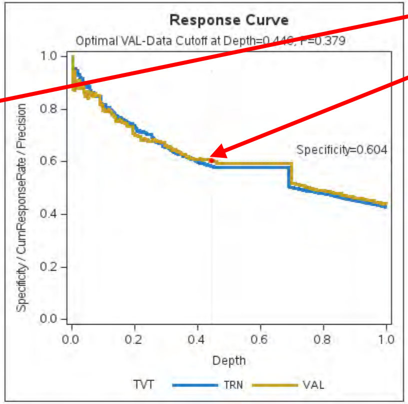
0. Step: VA TA

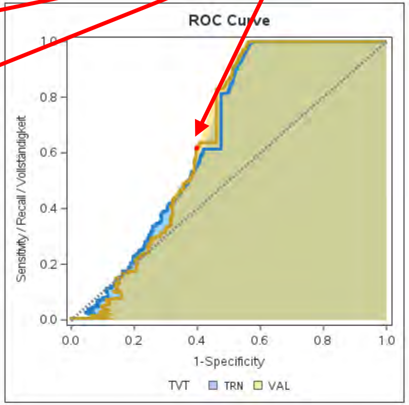
Corrupted Data

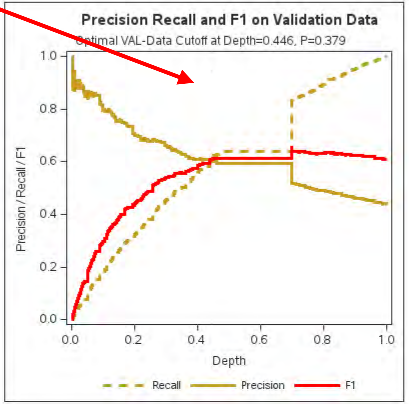
Assessment for Model 0.a Repaired Data (N=7575) Gradient Boosting with 6 Features { 6TextTopics}, Dependent Variable: Target

TVT	Depth	P_Cutoff	Sensitivity	Specificity	f1	lift	benefit	PriorProb	KS
TRN	0.36446	0.41863	0.52829	0.61853	0.56986	1.44953	0.16383	0.42671	0.28578
VAL	0.44635	0.37946	0.61732	0.60404	0.61061	1.38304	0.17097	0.43668	0.30350









Training Data, Cutoff used: 0.419

Frequency
Percent
Row Pct
Col Pct

Table 1 of target by Predicted
Controlling for TVT=TRN

target	Predicted 0	Predicted 1	Total
0	2302 43.43 75.75 68.31	737 13.90 24.25 38.17	3039 57.33
1	1068 20.15 47.21 31.69	1194 22.52 52.79 61.83	2262 42.67
Total	3370 63.57	1931 36.43	5301 100.00

Validation Data, Cutoff used: 0.379

Frequency
Percent
Row Pct
Col Pct

Table 1 of target by Predicted
Controlling for TVT=VAL

target	Predicted 0	Predicted 1	Total
0	879 38.65 68.62 69.82	402 17.68 31.38 39.61	1281 56.33
1	380 16.71 38.27 30.18	613 26.96 61.73 60.39	993 43.67
Total	1259 55.36	1015 44.64	2274 100.00

Selected Prediction Features: Top 11

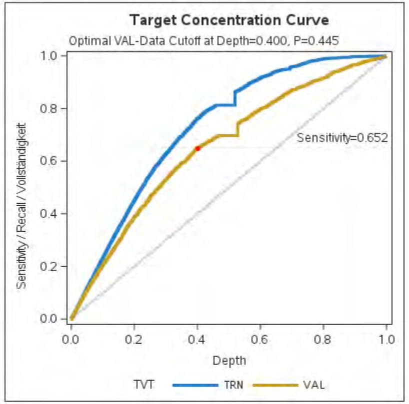
No	Variable Name and Label	RelImp	Number of Levels
1	_etm__Col1_: "not, +even, blood, +	1.00	2958
2	_etm__Col3_: "+wildfire, californi	0.78	754
3	_etm__Col5_: "+fire, +forest, +tru	0.71	2413
4	_etm__Col2_: "+detonate, army, +ol	0.63	1565
5	_etm__Col4_: "+confirm, mh370, wre	0.51	903
6	_etm__Col6_: "reddit, +quarantine,	0.25	2176

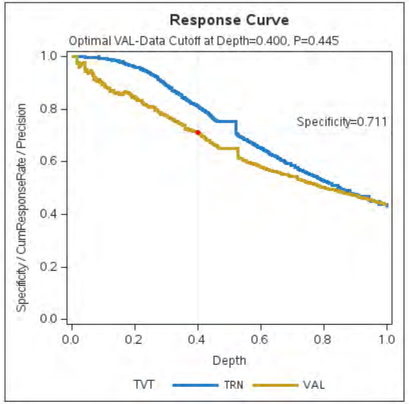
1. Step: VTA MS Pipeline

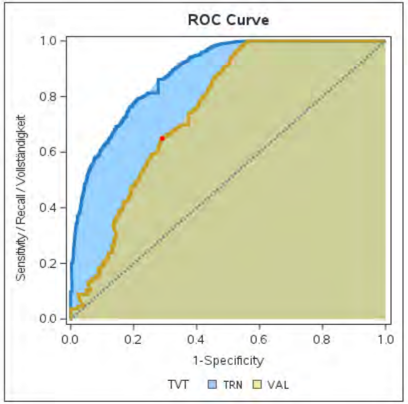
Corrupted Data

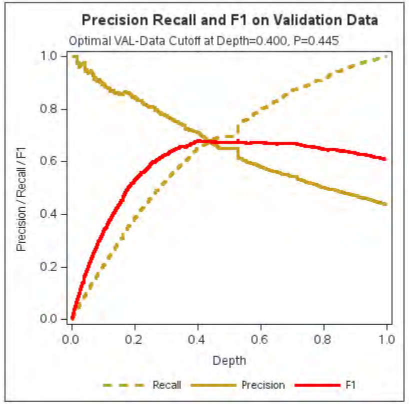
Assessment for Model 1. Gradient Boosting with 25 Features (25TextTopics), Dependent Variable: Target

TVT	Depth	P_Cutoff	Sensitivity	Specificity	f1	lift	benefit	PriorProb	KS
TRN	0.41873	0.42033	0.78214	0.79591	0.78896	1.86791	0.36342	0.42610	0.63323
VAL	0.40028	0.44498	0.65180	0.71147	0.68033	1.62838	0.25153	0.43692	0.44670









Training Data, Cutoff used: 0.420

Frequency
Percent
Row Pct
Col Pct

Table 1 of target by Predicted
Controlling for TVT=TRN

target	Predicted 0	Predicted 1	Total
0	2449 48.78 85.01 84.01	432 8.61 14.99 20.52	2881 57.39
1	466 9.28 21.79 15.99	1673 33.33 78.21 79.48	2139 42.61
Total	2915 58.07	2105 41.93	5020 100.00

Validation Data, Cutoff used: 0.445

Frequency
Percent
Row Pct
Col Pct

Table 1 of target by Predicted
Controlling for TVT=VAL

target	Predicted 0	Predicted 1	Total
0	965 44.76 79.49 74.57	249 11.55 20.51 28.89	1214 56.31
1	329 15.26 34.93 25.43	613 28.43 65.07 71.11	942 43.69
Total	1294 60.02	862 39.98	2156 100.00

Selected Prediction Features: Top 11

No	Variable Name and Label	RelImp	Number of Levels
1	_etm_COL5: "+fire, +forest, +truck	1.00	2353
2	_etm_COL23: "+go, +flame, +siren,	0.91	2660
3	_etm_COL10: "+bomb, hiroshima, ato	0.91	1897
4	_etm_COL18: "+scream, im, internal	0.86	1242
5	_etm_COL21: "+train, +life, +derai	0.84	2012
6	_etm_COL2: "+wildfire, california,	0.74	679
7	_etm_COL22: "still, +war, +world,	0.61	3211
8	_etm_COL7: "+bag, +body, +cross, +	0.56	1857
9	_etm_COL14: "+man, +car, +flame, +	0.54	3012
10	_etm_COL1: "amp, rt, +please, +bac	0.52	3473
11	_etm_COL4: "+confirm, wreckage, mh	0.48	473

Selected Prediction Features: Top 12-22

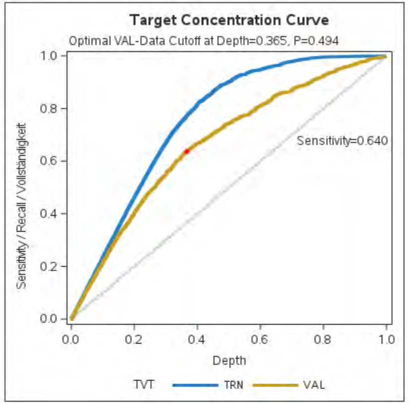
No	Variable Name and Label	RelImp	Number of Levels
12	_etm_COL19: "+wreck, +word, +stock	0.40	1424
13	_etm_COL25: "+migrant, +rescuer, h	0.40	1729
14	_etm_COL17: "+video, youtube, play	0.36	1690
15	_etm_COL16: "+wave, +hot, +hijack,	0.34	1553
16	_etm_COL20: "+crush, +woman, +girl	0.32	1253
17	_etm_COL9: "+legionnaire, +family,	0.30	385
18	_etm_COL6: "reddit, +quarantine, c	0.27	2152
19	_etm_COL13: "+burn, +build, not, +	0.25	1912
20	_etm_COL3: "+detonate, army, +old,	0.22	1176
21	_etm_COL11: "+disaster, obama, typ	0.15	888
22	_etm_COL15: "+oil, +spill, +big, +	0.15	509

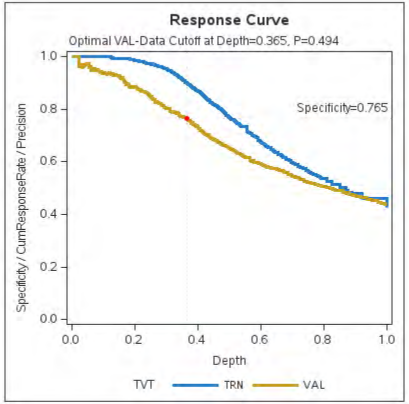
2. Step: VTA MS Pipeline

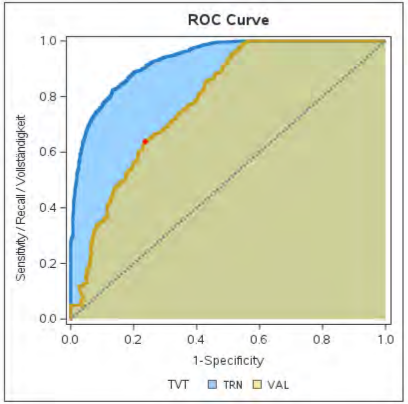
Corrupted Data

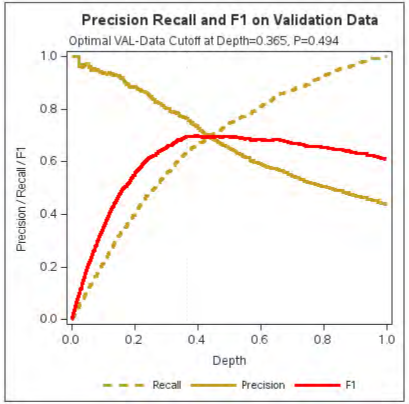
Assessment for Model 2. Gradient Boosting with 76 Features (66TextTopics), Dependent Variable: Target

TVT	Depth	P_Cutoff	Sensitivity	Specificity	f1	lift	benefit	PriorProb	KS
TRN	0.40498	0.44398	0.82328	0.86621	0.84420	2.03289	0.41830	0.42610	0.72887
VAL	0.36549	0.49405	0.64013	0.76523	0.69711	1.75141	0.27464	0.43692	0.48774









Training Data, Cutoff used: 0.444

Frequency
Percent
Row Pct
Col Pct

Table 1 of target by Predicted
Controlling for TVT=TRN

target	Predicted 0	Predicted 1	Total
0	2609 51.97 90.56 87.32	272 5.42 9.44 13.39	2881 57.39
1	379 7.55 17.72 12.68	1760 35.06 82.28 86.61	2139 42.61
Total	2988 59.52	2032 40.48	5020 100.00

Validation Data, Cutoff used: 0.494

Frequency
Percent
Row Pct
Col Pct

Table 1 of target by Predicted
Controlling for TVT=VAL

target	Predicted 0	Predicted 1	Total
0	1029 47.73 84.76 75.22	185 8.58 15.24 23.48	1214 56.31
1	339 15.72 35.99 24.78	603 27.97 64.01 76.52	942 43.69
Total	1368 63.45	788 36.55	2156 100.00

Selected Prediction Features: Top 11

No	Variable Name and Label	RelImp	Number of Levels
1	_etm_COL47: "+thunderstorm, severe	1.00	2496
2	_etm_COL5: "+fire, +forest, +truck	0.91	2353
3	_etm_COL66: "+good, +know, +way, +	0.88	4066
4	_etm_COL10: "+bomb, hiroshima, ato	0.87	1897
5	_etm_COL2: "+wildfire, california,	0.84	679
6	_etm_COL18: "+scream, im, internal	0.72	1242
7	_etm_COL36: "+suicide, +kill, saud	0.65	1440
8	_etm_COL59: "+new, +collide, full,	0.58	4075
9	_etm_COL21: "+train, +life, +derai	0.57	2012
10	_etm_COL72: "+charge, +boy, mansla	0.53	3352
11	_etm_COL23: "+go, +flame, +siren,	0.46	2660

Selected Prediction Features: Top 12-22

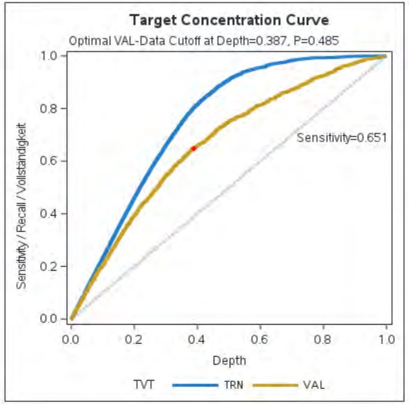
No	Variable Name and Label	RelImp	Number of Levels
12	_etm_COL56: "+day, +riot, +ruin, r	0.43	4244
13	_etm_COL71: "+see, +panic, +attack	0.43	3621
14	_etm_COL60: "+blow, +electrocute,	0.41	3469
15	_etm_COL46: "+people, +panic, +smo	0.40	4020
16	_etm_COL31: "not, +even, blood, +f	0.40	2854
17	_etm_COL33: "+accident, airplane,	0.40	2584
18	_etm_COL28: "police, +wound, +susp	0.36	1534
19	_etm_COL4: "+confirm, wreckage, mh	0.33	473
20	_etm_COL62: "+come, +smoke, here,	0.32	3790
21	_etm_COL27: "+live, +see, tragedy,	0.30	3102
22	_etm_COL53: "+fear, +ambulance, +h	0.30	1918

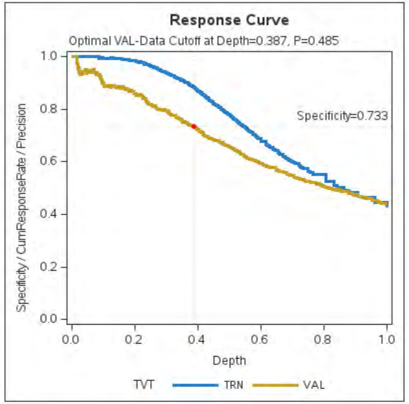
2a. Step: VTA MS Pipeline

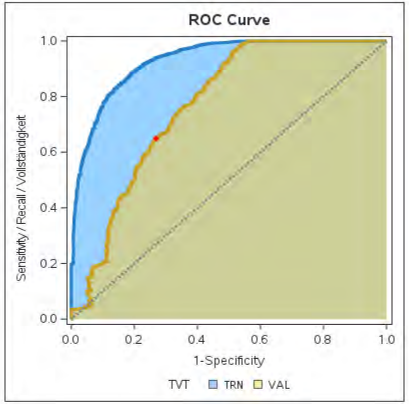
Clean Data

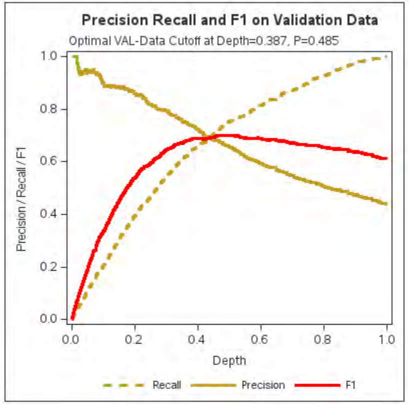
Model 2.a Repaired Data (N=7575): Gradient Boosting with 76 Features { 76TextTopics}, Dependent Variable: Target

TVT	Depth	P_Cutoff	Sensitivity	Specificity	f1	lift	benefit	PriorProb	KS
TRN	0.43105	0.43615	0.85013	0.84158	0.84583	1.97223	0.41908	0.42671	0.73101
VAL	0.38742	0.48493	0.65055	0.73326	0.68943	1.67918	0.26313	0.43668	0.46710 (prev. 0.48774)









Training Data, Cutoff used: 0.436

Frequency
Percent
Row Pct
Col Pct

Table 1 of target by Predicted
Controlling for TVT=TRN

target	Predicted 0	Predicted 1	Total
0	2677 50.50 88.09 88.76	362 6.83 11.91 15.84	3039 57.33
1	339 6.40 14.99 11.24	1923 36.28 85.01 84.16	2262 42.67
Total	3016 56.89	2285 43.11	5301 100.00

Validation Data, Cutoff used: 0.485

Frequency
Percent
Row Pct
Col Pct

Table 1 of target by Predicted
Controlling for TVT=VAL

target	Predicted 0	Predicted 1	Total
0	1046 46.00 81.65 75.04	235 10.33 18.35 26.70	1281 56.33
1	348 15.30 35.05 24.96	645 28.36 64.95 73.30	993 43.67
Total	1394 61.30	880 38.70	2274 100.00

Selected Prediction Features: Top 11

No	Variable Name and Label	RelImp	Number of Levels
1	_etm__Col51_: "+thunderstorm, seve	1.00	2539
2	_etm__Col28_: "+suicide, +kill, sa	0.85	1608
3	_etm__Col5_: "+fire, +forest, +tru	0.76	2413
4	_etm__Col67_: "+blow, +time, +elec	0.73	4284
5	_etm__Col44_: "ª, +let, +want, do	0.73	3569
6	_etm__Col24_: "+train, +life, dera	0.72	2436
7	_etm__Col10_: "+bomb, hiroshima, a	0.65	1886
8	_etm__Col9_: "+scream, im, arianag	0.64	967
9	_etm__Col62_: "+new, +collide, +we	0.63	4553
10	_etm__Col38_: "+love, +collide, +y	0.62	2644
11	_etm__Col73_: "+fuck, +back, +weap	0.60	3781

Selected Prediction Features: Top 12-22

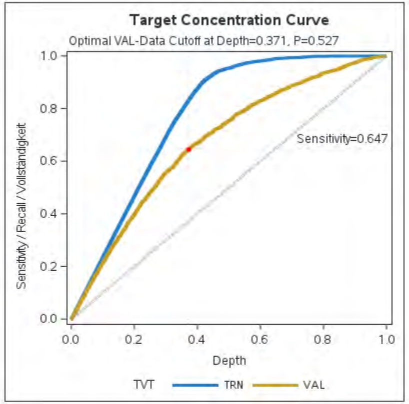
No	Variable Name and Label	RelImp	Number of Levels
12	_etm__Col22_: "police, +wind, +sus	0.59	1862
13	_etm__Col30_: "now, +right, +panic	0.58	3920
14	_etm__Col52_: "+day, +riot, +good,	0.56	4445
15	_etm__Col66_: "+flood, +work, +rai	0.55	3362
16	_etm__Col1_: "not, +even, blood, +	0.55	2958
17	_etm__Col29_: "emergency, +plan, +	0.50	3217
18	_etm__Col7_: "+bag, +body, +cross,	0.49	1636
19	_etm__Col32_: "+see, +back, +life,	0.49	4382
20	_etm__Col76_: "+charge, +boy, mans	0.46	4026
21	_etm__Col3_: "+wildfire, californi	0.45	754
22	_etm__Col25_: "+go, +siren, +let,	0.45	3313

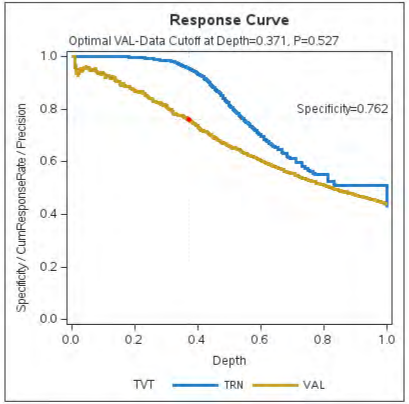
3a. Step: 147 Features

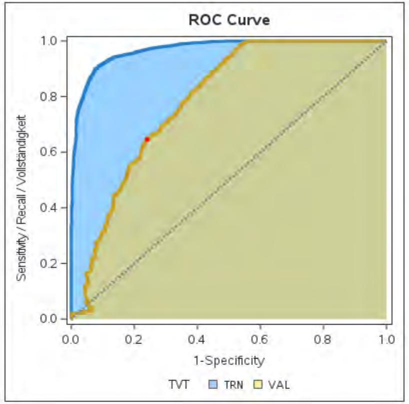
Clean Data

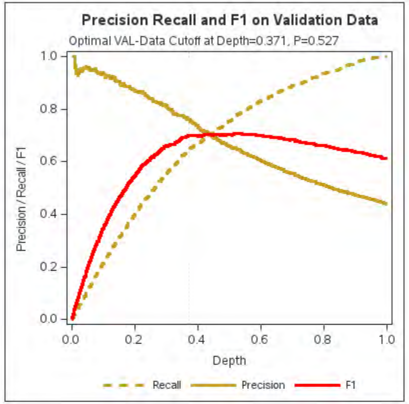
3.a Repaired Data (N=7575): Gradient Boosting with 147 Features {#Terms @Terms URLs WordCnt CharCnt 76TextTopics}, Dependent Variable: Target

TVT	Depth	P_Cutoff	Sensitivity	Specificity	f1	lift	benefit	PriorProb	KS
TRN	0.42501	0.43253	0.91026	0.91389	0.91207	2.14171	0.48524	0.42671	0.84642
VAL	0.37071	0.52697	0.64653	0.76157	0.69935	1.74401	0.27581	0.43668	0.48962









„Our new features“ not provided automatically in VA or VTA

Training Data, Cutoff used: 0.433

Frequency
Percent
Row Pct
Col Pct

Table 1 of target by Predicted
Controlling for TVT=TRN

target	Predicted 0	Predicted 1	Total
0	2845 53.67 93.62 93.25	194 3.66 6.38 8.62	3039 57.33
1	206 3.89 9.11 6.75	2056 38.79 90.89 91.38	2262 42.67
Total	3051 57.56	2250 42.44	5301 100.00

Validation Data, Cutoff used: 0.527

Frequency
Percent
Row Pct
Col Pct

Table 1 of target by Predicted
Controlling for TVT=VAL

target	Predicted 0	Predicted 1	Total
0	1080 47.49 84.31 75.42	201 8.84 15.69 23.87	1281 56.33
1	352 15.48 35.45 24.58	641 28.19 64.55 76.13	993 43.67
Total	1432 62.97	842 37.03	2274 100.00

Selected Prediction Features: Top 11

No	Variable Name and Label	RelImp	Number of Levels
1	_etm_N_Links: Number of Links in th	1.00	5
2	_etm__Col28_: "+suicide, +kill, sa	0.64	1608
3	_etm__Col51_: "+thunderstorm, seve	0.54	2539
4	_etm_chrctr_cnt: Number of Characte	0.51	176
5	_etm__Col5_: "+fire, +forest, +tru	0.49	2413
6	_etm__Col3_: "+wildfire, californi	0.46	754
7	_etm__Col24_: "+train, +life, dera	0.38	2436
8	_etm__Col10_: "+bomb, hiroshima, a	0.38	1886
9	_etm__Col22_: "police, +wind, +sus	0.37	1862
10	_etm__Col34_: "+accident, +airplan	0.34	3202
11	_etm__Col7_: "+bag, +body, +cross,	0.33	1636

Selected Prediction Features: Top 12-22

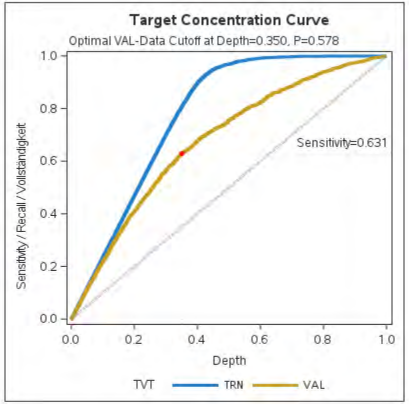
No	Variable Name and Label	RelImp	Number of Levels
12	_etm__Col75_: "+know, +good, +let,	0.32	4187
13	_etm__Col30_: "now, +right, +panic	0.30	3920
14	_etm__Col52_: "+day, +riot, +good,	0.30	4445
15	_etm__Col70_: "+year, ¥, +time, f	0.29	4384
16	_etm__Col38_: "+love, +collide, +y	0.27	2644
17	_etm__Col36_: "+mass, +murderer, +	0.26	2455
18	_etm_AtSign_3: 3. user mention in t	0.26	118
19	_etm__Col60_: "+rescue, +hostage,	0.26	3400
20	_etm__Col73_: "+fuck, +back, +weap	0.25	3781
21	_etm__Col67_: "+blow, +time, +elec	0.25	4284
22	_etm__Col44_: "", +let, +want, do	0.24	3569

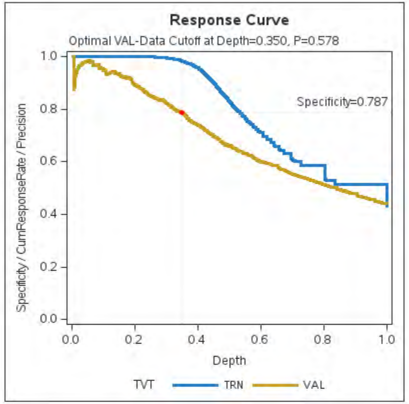
4a. Step: 149 Features

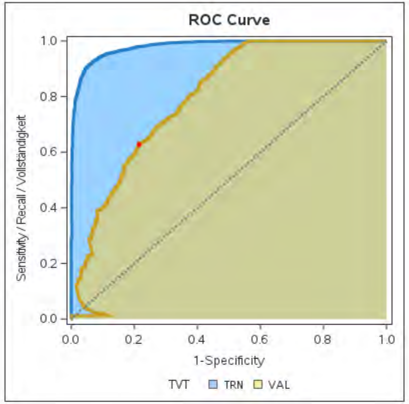
Clean Data

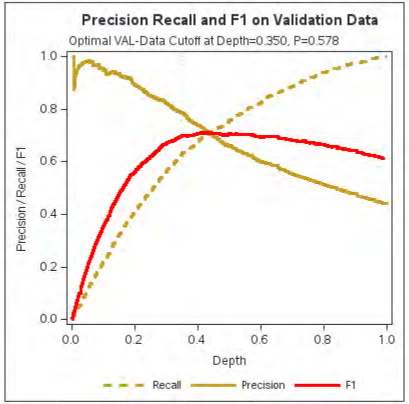
Repaired Data (N=7575): Gradient Boosting with 149 Features {#Terms @Terms URLs WordCnt CharCnt Sentiment 76TextTopics}, Dependent Variable: Target

TVT	Depth	P_Cutoff	Sensitivity	Specificity	f1	lift	benefit	PriorProb	KS
TRN	0.42558	0.43233	0.92750	0.92996	0.92873	2.17937	0.50192	0.42671	0.87551
VAL	0.35048	0.57797	0.63142	0.78670	0.70056	1.80157	0.28094	0.43668	0.49871









Training Data, Cutoff used: 0.432

Frequency
Percent
Row Pct
Col Pct

Table 1 of target by Predicted
Controlling for TVT=TRN

target	Predicted 0	Predicted 1	Total
0	2881 54.35 94.80 94.61	158 2.98 5.20 7.00	3039 57.33
1	164 3.09 7.25 5.39	2098 39.58 92.75 93.00	2262 42.67
Total	3045 57.44	2256 42.56	5301 100.00

Validation Data, Cutoff used: 0.578

Frequency
Percent
Row Pct
Col Pct

Table 1 of target by Predicted
Controlling for TVT=VAL

target	Predicted 0	Predicted 1	Total
0	1111 48.86 86.73 75.17	170 7.48 13.27 21.36	1281 56.33
1	367 16.14 36.96 24.83	626 27.53 63.04 78.64	993 43.67
Total	1478 65.00	796 35.00	2274 100.00

Selected Prediction Features: Top 11

No	Variable Name and Label	RelImp	Number of Levels
1	_etm_sentiment_score: Score value f	1.00	14
2	_etm_cooc_Link_4: Co-Occurence Link	0.82	2
3	_etm_prcntUsd: Percentage used of t	0.78	176
4	_etm__Col5_: "+fire, +forest, +tru	0.64	2413
5	_etm__Col3_: "+wildfire, californi	0.59	754
6	_etm__Col10_: "+bomb, hiroshima, a	0.58	1886
7	_etm__Col51_: "+thunderstorm, seve	0.53	2539
8	_etm__Col75_: "+know, +good, +let,	0.50	4187
9	_etm__Col24_: "+train, +life, dera	0.45	2436
10	_etm__Col22_: "police, +wind, +sus	0.45	1862
11	_etm_N_Links: Number of Links in th	0.43	5

Selected Prediction Features: Top 12-22

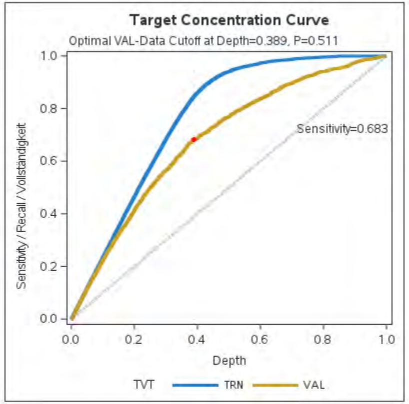
No	Variable Name and Label	RelImp	Number of Levels
12	_etm__Col30_: "now, +right, +panic	0.39	3920
13	_etm_chrctr_cnt: Number of Characte	0.38	176
14	_etm__Col52_: "+day, +riot, +good,	0.37	4445
15	_etm__Col7_: "+bag, +body, +cross,	0.36	1636
16	_etm__Col46_: "debris, +find, reun	0.35	2298
17	_etm__Col34_: "+accident, +airplan	0.34	3202
18	_etm__Col44_: "*, +let, +want, do	0.31	3569
19	_etm__Col9_: "+scream, im, arianag	0.30	967
20	_etm__Col73_: "+fuck, +back, +weap	0.29	3781
21	_etm__Col72_: "+say, +world, +elec	0.28	4730
22	_etm_wrd_cnt: Number of Words in a	0.28	31

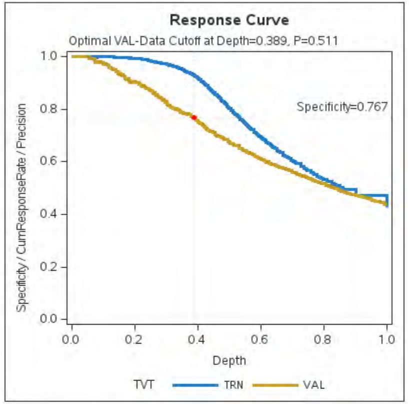
4a. Step: 160 Features

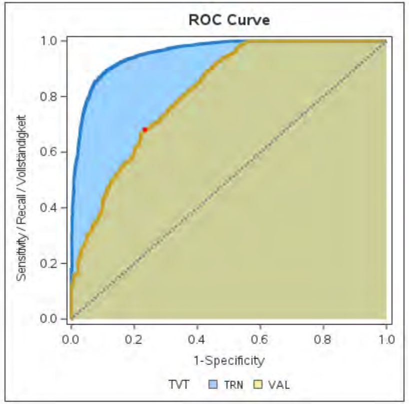
Clean Data

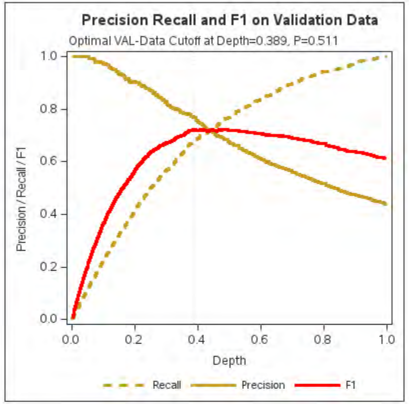
d Data (N=7575): Gradient Boosting with 160 Features (PreDefConcepts #Terms @Terms URLs WordCnt CharCnt Sentiment 76TextTopics), Dependent Variable

TVT	Depth	P_Cutoff	Sensitivity	Specificity	f1	lift	benefit	PriorProb	KS
TRN	0.42916	0.41800	0.89346	0.88835	0.89090	2.08185	0.46429	0.42671	0.80988
VAL	0.38874	0.51091	0.68278	0.76697	0.72243	1.75638	0.29404	0.43668	0.52197









Training Data, Cutoff used: 0.418

Frequency / Percent / Row Pct / Col Pct

Table 1 of target by Predicted, Controlling for TVT=TRN

target	Predicted 0	Predicted 1	Total
0	2785 52.54 91.64 92.04	254 4.79 8.36 11.16	3039 57.33
1	241 4.55 10.65 7.96	2021 38.12 89.35 88.84	2262 42.67
Total	3026 57.08	2275 42.92	5301 100.00

Validation Data, Cutoff used: 0.511

Frequency / Percent / Row Pct / Col Pct

Table 1 of target by Predicted, Controlling for TVT=VAL

target	Predicted 0	Predicted 1	Total
0	1075 47.27 83.92 77.28	206 9.06 16.08 23.33	1281 56.33
1	316 13.90 31.82 22.72	677 29.77 68.18 76.67	993 43.67
Total	1391 61.17	883 38.83	2274 100.00

Selected Prediction Features: Top 11

No	Variable Name and Label	RelImp	Number of Levels
1	_etm_concept: The concept that was	1.00	10
2	_etm_sentiment: Sentiment of the te	0.34	3
3	_etm_N_FullStop: Number of Periods	0.33	16
4	_etm_N_Links: Number of Links in th	0.23	5
5	_etm__Col51_: "+thunderstorm, seve	0.22	2539
6	_etm__Col5_: "+fire, +forest, +tru	0.22	2413
7	_etm__Col10_: "+bomb, hiroshima, a	0.15	1886
8	_etm__Col28_: "+suicide, +kill, sa	0.15	1608
9	_etm__Col7_: "+bag, +body, +cross,	0.15	1636
10	_etm__Col62_: "+new, +collide, +we	0.13	4553
11	_etm_chrctr_cnt: Number of Characte	0.13	176

Selected Prediction Features: Top 12-22

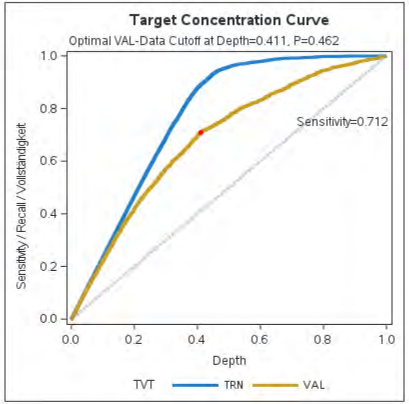
No	Variable Name and Label	RelImp	Number of Levels
12	_etm__Col44_: "*, +let, +want, do	0.13	3569
13	_etm_sentiment_score: Score value f	0.12	14
14	_etm__Col32_: "+see, +back, +life,	0.11	4382
15	_etm__Col67_: "+blow, +time, +elec	0.10	4284
16	_etm__Col24_: "+train, +life, dera	0.10	2436
17	_etm__Col73_: "+fuck, +back, +weap	0.10	3781
18	_etm_cooc_Link_4: Co-Occurence Link	0.09	2
19	_etm__Col75_: "+know, +good, +let,	0.09	4187
20	_etm_wrd_cnt: Number of Words in a	0.09	31
21	_etm_total_concepts: Total Number o	0.09	23
22	_etm__Col34_: "+accident, +airplan	0.09	3202

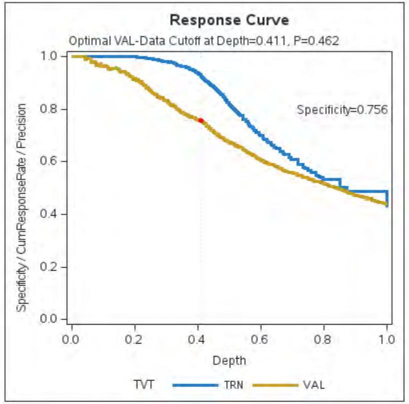
5a. Step: 326 Features

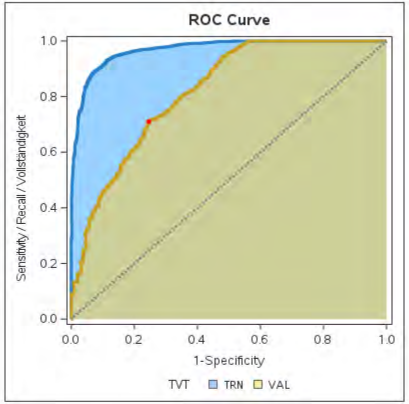
Clean Data

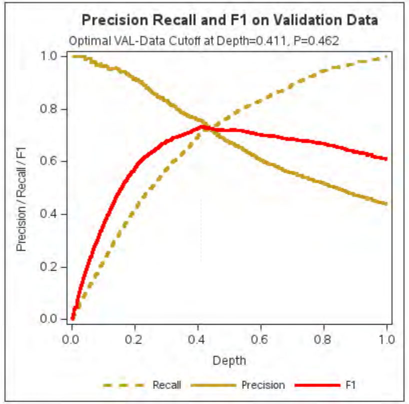
a (N=7575): Gradient Boosting with 326 Features {BoolRules PreDefConcepts #Terms @Terms URLs WordCnt CharCnt Sentiment 76TextTopics}, Dependent

TVT	Depth	P_Cutoff	Sensitivity	Specificity	f1	lift	benefit	PriorProb	KS
TRN	0.44709	0.37785	0.93015	0.88776	0.90846	2.08048	0.48306	0.42671	0.84262
VAL	0.41117	0.46210	0.71198	0.75615	0.73340	1.73161	0.30081	0.43668	0.53400









Training Data, Cutoff used: 0.378

Frequency
Percent
Row Pct
Col Pct

Table 1 of target by Predicted
Controlling for TVT=TRN

target	0	1	Total
0	2773 52.31 91.25 94.58	266 5.02 8.75 11.23	3039 57.33
1	159 3.00 7.03 5.42	2103 39.67 92.97 88.77	2262 42.67
Total	2932 55.31	2369 44.69	5301 100.00

Validation Data, Cutoff used: 0.462

Frequency
Percent
Row Pct
Col Pct

Table 1 of target by Predicted
Controlling for TVT=VAL

target	0	1	Total
0	1053 46.31 82.20 78.64	228 10.03 17.80 24.39	1281 56.33
1	286 12.58 28.80 21.36	707 31.09 71.20 75.61	993 43.67
Total	1339 58.88	935 41.12	2274 100.00

Selected Prediction Features: Top 11

No	Variable Name and Label	RelImp	Number of Levels
1	BR1_0: Presence of TG1 assoc. BR	1.00	2
2	_etm_concept: The concept that was	0.35	10
3	_etm_cooc_Link_4: Co-Occurence Link	0.20	2
4	BR1_20: BR1_20: mp	0.10	2
5	_etm_sentiment_score: Score value f	0.07	14
6	_etm_sentiment: Sentiment of the te	0.07	3
7	_etm__Col51_: "+thunderstorm, seve	0.06	2539
8	_etm__Col12_: "amp, rt, +please, c	0.06	3648
9	_etm_total_concepts: Total Number o	0.06	23
10	_etm__Col44_: "*, +let, +want, do	0.05	3569
11	_etm__Col10_: "+bomb, hiroshima, a	0.05	1886

Selected Prediction Features: Top 12-22

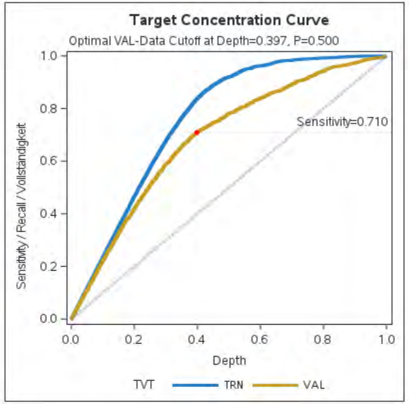
No	Variable Name and Label	RelImp	Number of Levels
12	_etm__Col9_: "+scream, im, arianag	0.05	967
13	_etm__Col5_: "+fire, +forest, +tru	0.05	2413
14	_etm_chrctr_cnt: Number of Characte	0.04	176
15	_etm__Col75_: "+know, +good, +let,	0.04	4187
16	_etm__Col62_: "+new, +collide, +we	0.04	4553
17	BR1_41: BR1_41: now	0.04	2
18	_etm__Col69_: "+say, +need, +stop,	0.04	4849
19	BR0_0: Presence of TG0 assoc. BR	0.04	2
20	_etm__Col68_: "+people, +panic, +s	0.04	4060
21	_etm__Col32_: "+see, +back, +life,	0.04	4382
22	_etm__Col28_: "+suicide, +kill, sa	0.04	1608

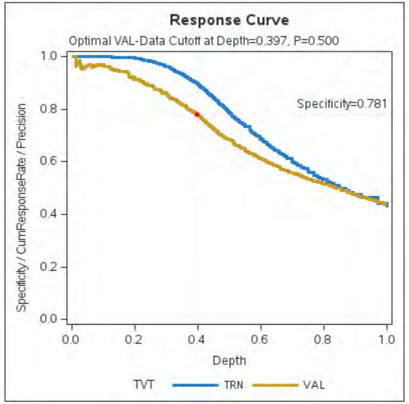
6a. Step: 550 Features

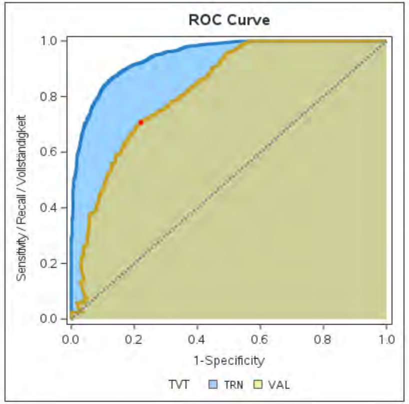
Clean Data

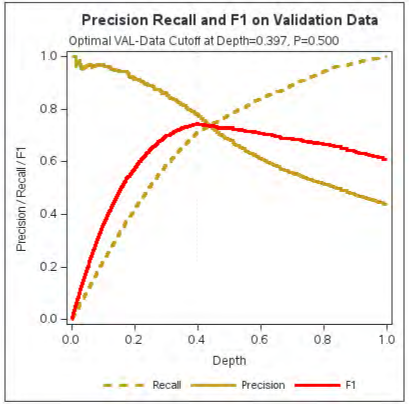
7575): Gradient Boosting with 550 Features {Keywords BooleRules PreDefConcepts #Terms @Terms URLs WordCnt CharCnt Sentiment 76TextTopics}, Deper

TVT	Depth	P_Cutoff	Sensitivity	Specificity	f1	lift	benefit	PriorProb	KS
TRN	0.41615	0.45210	0.85809	0.87987	0.86885	2.06198	0.44194	0.42671	0.77089
VAL	0.39710	0.49997	0.70997	0.78073	0.74367	1.78790	0.31287	0.43668	0.55540









Training Data, Cutoff used: 0.452

Frequency
Percent
Row Pct
Col Pct

Table 1 of target by Predicted
Controlling for TVT=TRN

target	Predicted 0	Predicted 1	Total
0	2774 52.33 91.28 89.63	265 5.00 8.72 12.01	3039 57.33
1	321 6.06 14.19 10.37	1941 36.62 85.81 87.99	2262 42.67
Total	3095 58.39	2206 41.61	5301 100.00

Validation Data, Cutoff used: 0.500

Frequency
Percent
Row Pct
Col Pct

Table 1 of target by Predicted
Controlling for TVT=VAL

target	Predicted 0	Predicted 1	Total
0	1083 47.63 84.54 78.94	198 8.71 15.46 21.95	1281 56.33
1	289 12.71 29.10 21.06	704 30.96 70.90 78.05	993 43.67
Total	1372 60.33	902 39.67	2274 100.00

Selected Prediction Features: Top 11

No	Variable Name and Label	RelImp	Number of Levels
1	KW1_C: Keyword TG1	1.00	80
2	KW0_C: Keyword TG0	0.94	107
3	BR1_0: Presence of TG1 assoc. BR	0.25	2
4	_etm_concept: The concept that was	0.19	10
5	_etm_N_Links: Number of Links in th	0.07	5
6	_etm_sentiment: Sentiment of the te	0.07	3
7	KW1_N: Number of Keywords TG1	0.04	7
8	_etm__Col44_: "°, +let, +want, do	0.04	3569
9	KW0_N: Number of Keywords TG0	0.04	6
10	_etm__Col51_: "+thunderstorm, seve	0.04	2539
11	_etm_wrd_cnt: Number of Words in a	0.03	31

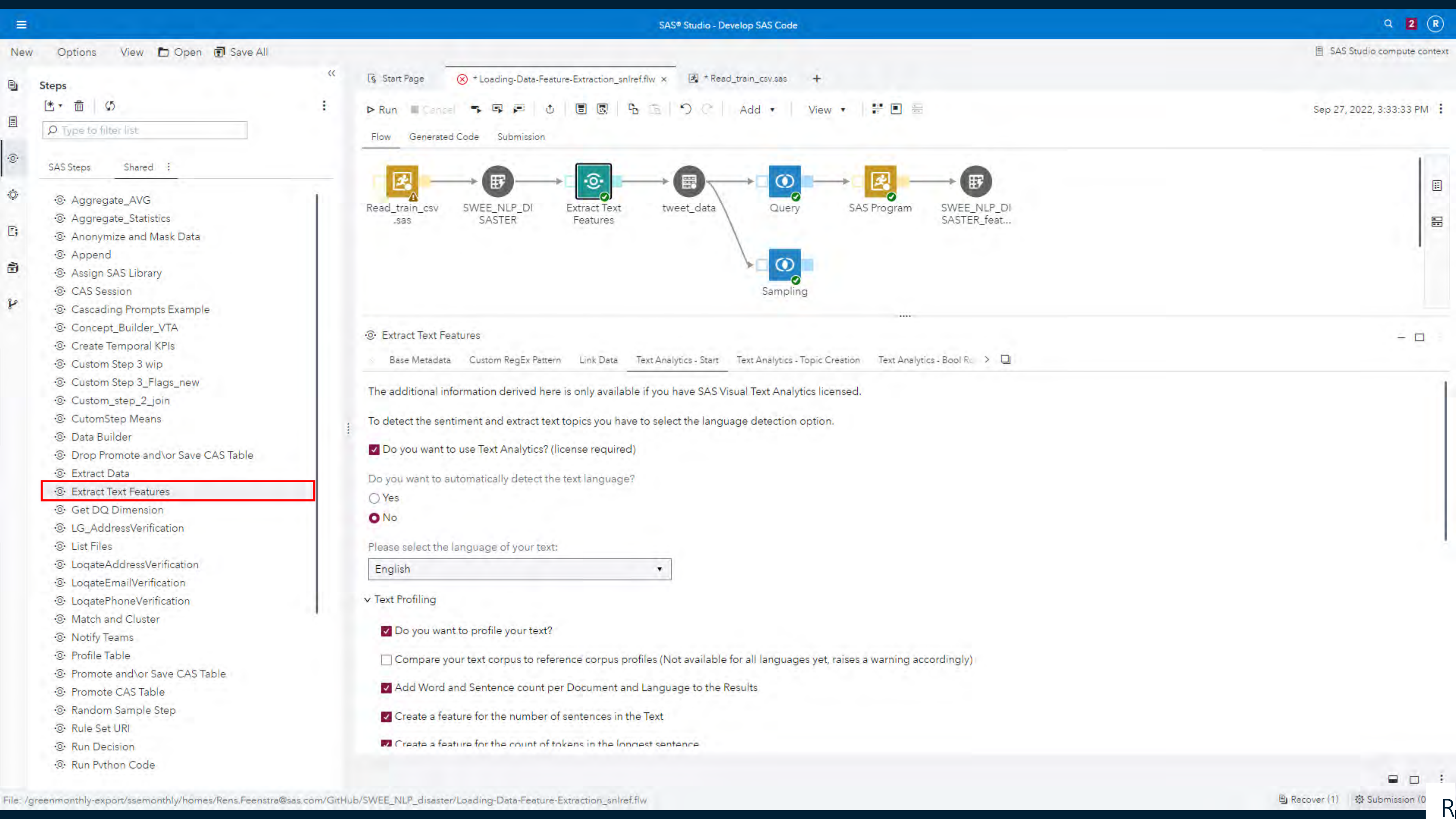
Selected Prediction Features: Top 12-22

No	Variable Name and Label	RelImp	Number of Levels
12	_etm__Col38_: "+love, +collide, +y	0.03	2644
13	BR1_62: BR1_62: fire	0.03	2
14	_etm__Col61_: "+make, +deluge, +ri	0.03	3843
15	_etm__Col24_: "+train, +life, dera	0.03	2436
16	_etm__Col71_: "+see, +live, traged	0.03	3420
17	_etm__Col29_: "emergency, +plan, +	0.03	3217
18	_etm_chrctr_cnt: Number of Characte	0.03	176
19	_etm__Col62_: "+new, +collide, +we	0.03	4553
20	_etm__Col70_: "+year, ¥, +time, f	0.03	4384
21	_etm__Col50_: "+school, hijacker,	0.03	2414
22	_etm_max_tokens_sentence: Number of	0.03	51

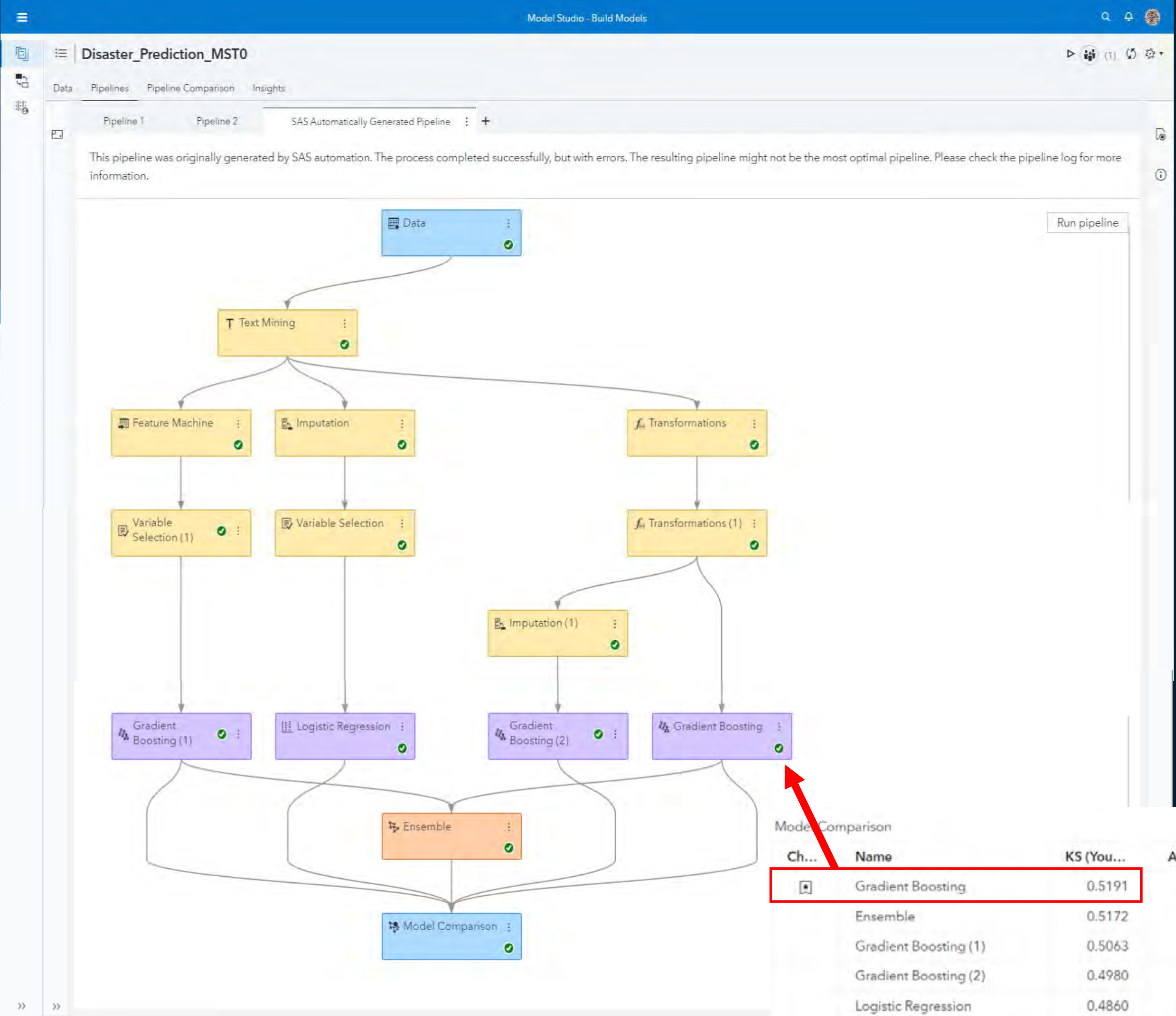
How can you use the custom Step

In SAS Studio on ssemonthly

- Modelling in ModelStudio

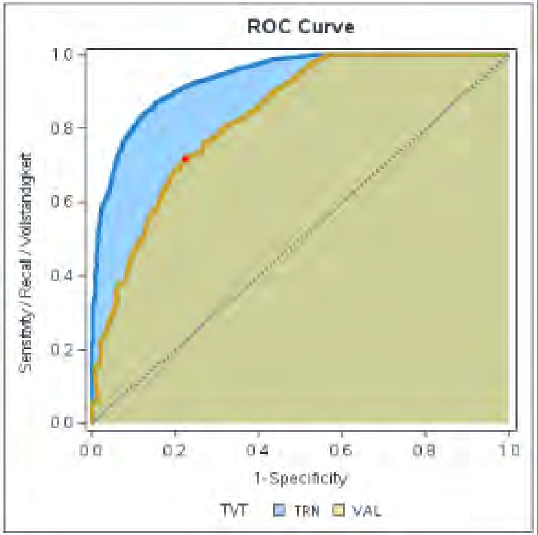


SAS® Studio - Develop SAS Code
New
Options
View
Open
Save All
SAS Studio compute context
Steps
Type to filter list
SAS Steps
Shared
Aggregate_AVG
Aggregate_Statistics
Anonymize and Mask Data
Append
Assign SAS Library
CAS Session
Cascading Prompts Example
Concept_Builder_VTA
Create Temporal KPIs
Custom Step 3 wip
Custom Step 3_Flags_new
Custom_step_2_join
CutomStep Means
Data Builder
Drop Promote and\or Save CAS Table
Extract Data
Extract Text Features
Get DQ Dimension
LG_AddressVerification
List Files
LoqateAddressVerification
LoqateEmailVerification
LoqatePhoneVerification
Match and Cluster
Notify Teams
Profile Table
Promote and\or Save CAS Table
Promote CAS Table
Random Sample Step
Rule Set URI
Run Decision
Run Python Code
Start Page
* Loading-Data-Feature-Extraction_snlref.flw
* Read_train_csv.sas
Run
Cancel
Add
View
Sep 27, 2022, 3:33:33 PM
Flow
Generated Code
Submission
Read_train_csv .sas
SWEE_NLP_DI SASTER
Extract Text Features
tweet_data
Query
SAS Program
SWEE_NLP_DI SASTER_feat...
Sampling
Extract Text Features
Base Metadata
Custom RegEx Pattern
Link Data
Text Analytics - Start
Text Analytics - Topic Creation
Text Analytics - Bool R
The additional information derived here is only available if you have SAS Visual Text Analytics licensed.
To detect the sentiment and extract text topics you have to select the language detection option.
Do you want to use Text Analytics? (license required)
Do you want to automatically detect the text language?
Yes
No
Please select the language of your text:
English
Text Profiling
Do you want to profile your text?
Compare your text corpus to reference corpus profiles (Not available for all languages yet, raises a warning accordingly)
Add Word and Sentence count per Document and Language to the Results
Create a feature for the number of sentences in the Text
Create a feature for the count of tokens in the longest sentence
File: /greenmonthly-export/ssemonthly/homes/Rens.Feenstra@sas.com/GitHub/SWEE_NLP_disaster/Loading-Data-Feature-Extraction_snlref.flw
Recover (1)
Submission (0



Model Comparison

Ch...	Name	KS (You...	Accuracy	Averag...	Area Un...	Cumula...	Cumula...	Cutoff	Data Role
★	Gradient Boosting	0.5191	0.7723	0.1669	0.8164	2.1762	21.7622	0.5000	VALIDATE
	Ensemble	0.5172	0.7709	0.1634	0.8221	2.1762	21.7622	0.5000	VALIDATE
	Gradient Boosting (1)	0.5063	0.7667	0.1666	0.8170	2.1868	21.8684	0.5000	VALIDATE
	Gradient Boosting (2)	0.4980	0.7630	0.1667	0.8162	2.2187	22.1868	0.5000	VALIDATE
	Logistic Regression	0.4860	0.7463	0.1742	0.8031	2.1762	21.7622	0.5000	VALIDATE



Result:

New Custom Step for SAS Studio

Creates ~550 prediction features

~ 10 feature request,

~detecting 12 bugs ,

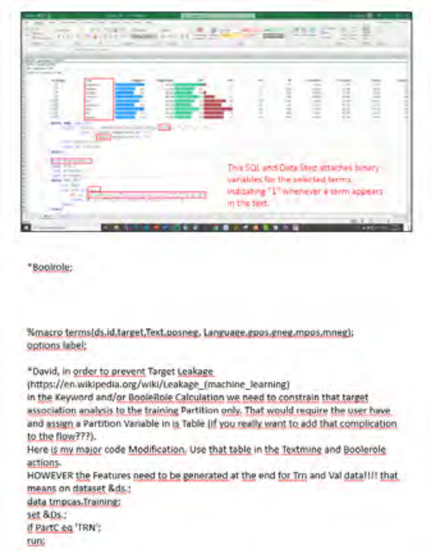
25 commits

…btw. you'll find the relevant SAS code snipplets in the notes below the slides of this slide deck

regex

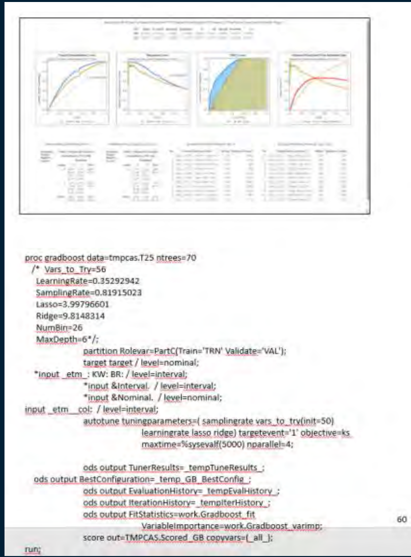

```
data test;
set mysasfil._ulr_tweets_desaster_pred_kaggle;
run;

*Required Input Parameters;
%let ds=Test;
%let Prx=/accident|crash|lost|drug|rockn|roll|fire|flood|outbreak|bag/i;/\b[a-
z]{2,3}\s{0,1}\d{2,4}\b/i;*/[a-z]{2,3}\s{0,1}\d{2,4\}\b/i; *Airlin Flug code  z.B. MH370;
%let ID=id;
%let Text=Text;
%let MinDocCNT=10;
%let Part=PartN;
%let TRN=(PartN eq 1);*Logic condition to subset for Training Data;

*Optional Input Parameters;

%let target=target;*Required to be a niumerical Binary variable with values 0 or 1;
%let MinTgMean=.8;
%let MaxTgMean=.2;

*Optional System Parameter (They must hace values!!! 10 and 32 ist  Default;
%let Snipplet_Window=10;*Number of characters extracte before and after
Orrurance;
%let Len_Occur=32;
```

Boolerule




```
*Boolrole;



%macro terms(ds,id,target,Text,posneg, Language,gpos,gneg,mpos,mneg);
options label;

*David, in order to prevent Target Leakage
(https://en.wikipedia.org/wiki/Leakage_(machine_learning)
in the Keyword and/or BooleRole Calculation we need to constrain that target
association analysis to the training Partition only. That would require the user have
and assign a Partition Variable in is Table (if you really want to add that complication
to the flow???).
Here is my major code Modification. Use that table in the Textmine and Boolerole
actions.
HOWEVER the Features need to be generated at the end for Trn and Val data!!!! that
means on dataset &ds.;
data tmpcas.Training;
set &Ds.;
if PartC eq 'TRN';
run;
```

Gradboost Output

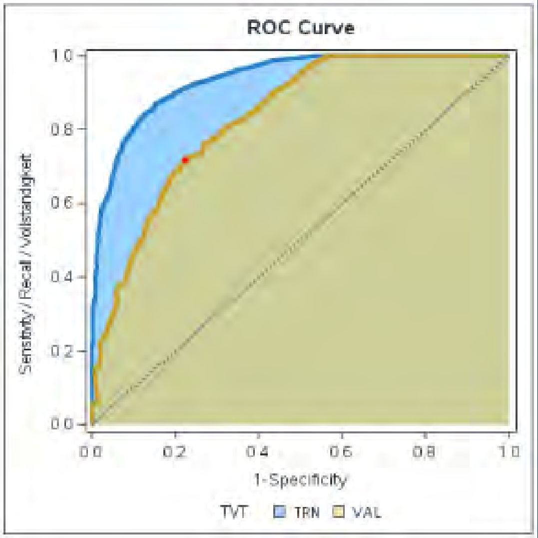


```
proc gradboost data=tmpcas.T25 ntrees=70
  /*  Vars_to_Try=56
    LearningRate=0.35292942
    SamplingRate=0.81915023
    Lasso=3.99796601
    Ridge=9.8148314
    NumBin=26
    MaxDepth=6*/;
                partition Rolevar=PartC(Train='TRN' Validate='VAL');
                target target / level=nominal;
   *input _etm_: KW: BR: / level=interval;
                *input &Interval.  / level=interval;
                *input &Nominal.  / level=nominal;
input _etm__col: / level=interval;
                autotune tuningparameters=( samplingrate vars_to_try(init=50)
                                learningrate lasso ridge) targetevent='1' objective=ks
                                maxtime=%sysevalf(5000) nparallel=4;

                ods output TunerResults=_tempTuneResults_;
   ods output BestConfiguration=_temp_GB_BestConfig_;
                ods output EvaluationHistory=_tempEvalHistory_;
                ods output IterationHistory=_tempIterHistory_;
                ods output FitStatistics=work.Gradboost_fit
                                VariableImportance=work.Gradboost_varimp;
                score out=TMPCAS.Scored_GB copyvars=(_all_);
run;
```
60

sas.com

Weiterführende Literatur

- SAS® Visual Text Analytics 8.5: User's Guide (Viya)
- SAS® Text Miner 14.1 Reference Help (V9)
- SAS® Text Analytics for Business Applications: Concept Rules for Information Extraction Models
- Git Repository von David Weik für den Text Analytics Flow (Nutzung auf eigene Gefahr!!!)



Questions
on *“climbing the ROC”*

Ulrich Reincke & David Weik