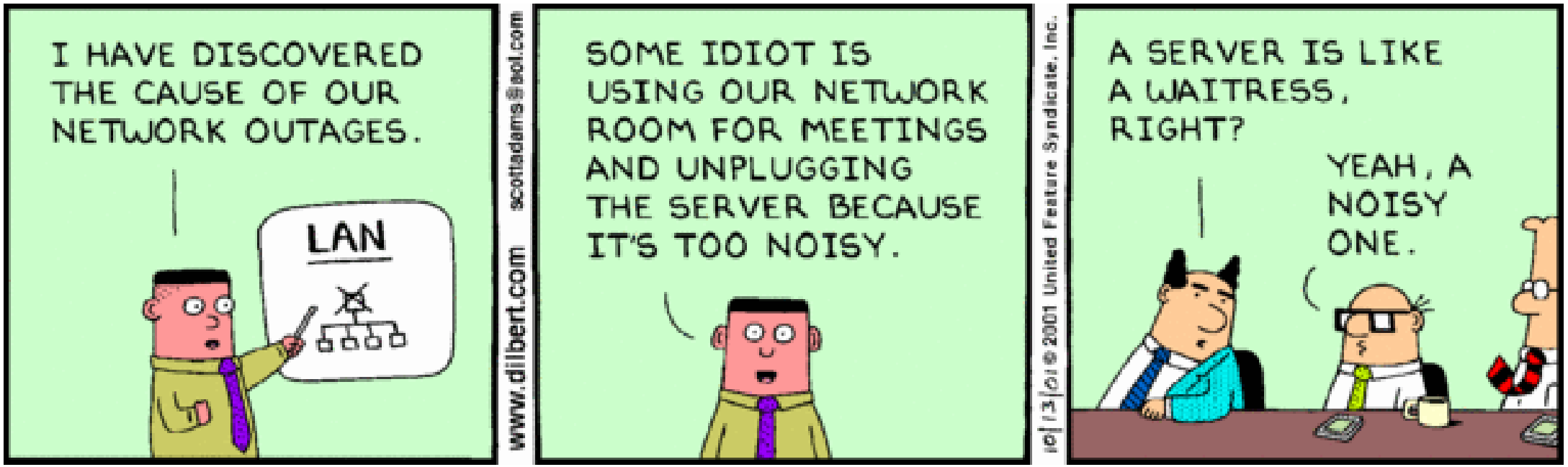
Capstone Project

Prediction of outages using Network Logs

May 2018

Rakesh Auplish

“The Problem”



Source: Dilbert.com

Background

- A telecom company is on a journey to enhance customer experience and putting customer first.
- Top priority items are network uptime and network quality of service.
- The aspiration is to be proactive in network maintenance.
- A statistical model for predicting network outages and their severity is required to facilitate proactive maintenance.

Exploratory Data Analysis

Summary

- 18552 Unique Log IDs together in – Test and Train
- 18552 Unique Log IDs in all other Data sets.
  All logs are included in data sets
- One or more rows for each Log ID in the data sets
- Mutually exclusive Log IDs across Test and Train
  0 rows repeat
- Location IDs repeat across Test and Train
  842 rows repeat

Data Frame	Rows	Columns	Unique IDs
Train	7381	3	7381
Test	11171	2	11171
Event	31170	2	18552
Resource	21076	2	18552
Severity	18552	2	18552
Log Feature	58671	3	18552

Data relationships

Logs (train model)		
id	location	fault_severity
14804	location 120	0

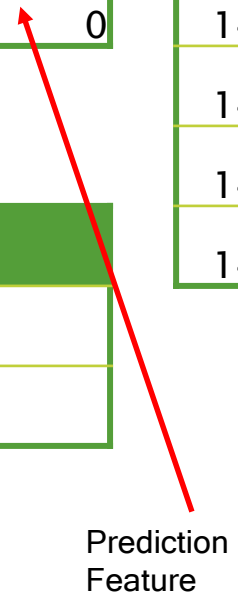
Logs (test model)		
id	location	
11066	location 481	

Event type	
id	event_type
14804	event_type 34
14804	event_type 11
14804	event_type 36
14804	event_type 20

Resource type	
id	resource_type
14804	resource_type 2
14804	resource_type 8

Severity type	
id	severity_type
14804	severity_type 1

Log feature		
id	log_feature	volume
14804	feature 134	1
14804	feature 219	1
14804	feature 117	1
14804	feature 227	2
14804	feature 237	2
14804	feature 232	2
14804	feature 181	1
14804	feature 160	1
14804	feature 29	1



Executive Summary

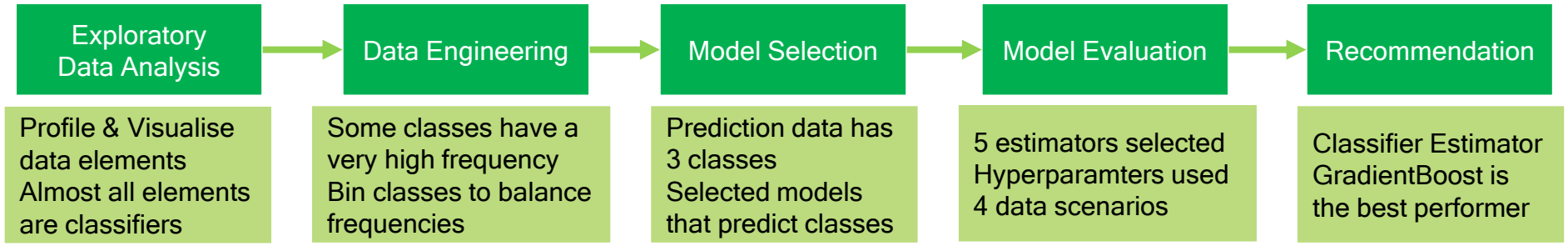
Objective

- Prediction of 3 class outcome of an incident on the telephone network. Develop a machine learning model to perform automatic predictions.
  - Possible outcomes – 0: No disruption, 1: Momentary glitch, 2: Service disruption

Features / Metrics

- 5 features (input metrics) used for prediction
  - Location, Logfeature and Volume, Event Type, Resource, Incident Severity

Methodology



Assumptions

- Data definitions and business meaning of data assumed
  - No metadata / data dictionary provided
- Relationships between data entities assumed from profiling data

Risks

- Assumptions about business meaning could be wrong
- An incorrect inference of relationship between data entities could have a profound impact on prediction accuracy

Results

- Baseline Accuracy 64.82% computed on provided data
- GradientBoost Classifier is the best estimator in all scenarios
  - 76.48% accuracy with hyperparameters 76.16% accuracy with default parameters
- Best performance time ~ average 32 sec per iteration
- Top 20 metrics influencing the predictions

	Metric	Coefficient		Metric	Coefficient
1	log_feature203	0.124773	11	log_feature179	0.016276
2	log_feature170	0.034011	12	severity_type_1	0.014595
3	log_feature202	0.03206	13	log_feature134	0.01443
4	log_feature209	0.024589	14	log_feature315	0.014178
5	log_feature232	0.024231	15	log_feature70	0.014025
6	log_feature312	0.023538	16	log_feature368	0.013464
7	log_feature73	0.023496	17	log_feature227	0.012689
8	log_feature82	0.018607	18	log_feature314	0.012604
9	log_feature171	0.018412	19	log_feature54	0.012336
10	log_feature155	0.016335	20	event_type_OTH	0.012102

Recommendation

- GradientBoost Classifier model
  - has consistently given best accuracy results
  - has shown the best performance with average execution time of ~32 secs
- The model will be suitable for machine learning implementation
- The model will also be suitable for real-time predictions due to its fast execution

Questions

Source: Dilbert.com

Appendixes

Data Dictionary

Data Dictionary is based on inferences drawn from data profiling and analysis

Entity	Logs	
Description	Periodic information provided by equipment sensors connected to the network relating to state of network	
Attributes	Id	Unique Identifier for a log
	Location	Id for the location of the equipment (no description)
	Fault Severity	Resultant severity
	3 classes *(0 - No disruption, 1 - Momentary glitch, 2 - Total disruption)*	
Comments	Two data sets have been provided	
	train	*dataset for training the models, contains fault_severity*
	test	*dataset for testing the models*

Data Dictionary

Data Dictionary is based on inferences drawn from data profiling and analysis

Entity	Events	
Description	Type of event reported by equipment sensors. Multiple event types can be associated to a Log	
Attributes	Id	Unique Identifier for a log
	Event_type	Id for event type(no description)
	Multiple Classes - values like 11, 15, 20, 7	
Comment	One data set has been provided	

Data Dictionary

Data Dictionary is based on inferences drawn from data profiling and analysis

Entity	Resources	
Description	Type of resource providing information to a log. Multiple resources can be associated to a Log	
Attributes	Id	Unique Identifier for a log
	resource_type	Id for a resource (no description)
		Multiple classes - v*alues: 1,2,3,4,5,6,7,8,9,10*
Comment	One data set has been provided	

Data Dictionary

Data Dictionary is based on inferences drawn from data profiling and analysis

Entity	Severity Type	
Description	Type of severity for the log. One severity associated to a Log	
Attributes	Id	Unique Identifier for a log
	severity_type	Id for the severity (no description)
	Multiple classes - values: 1, 2, 3, 4, 5	
Comment	One data set has been provided	

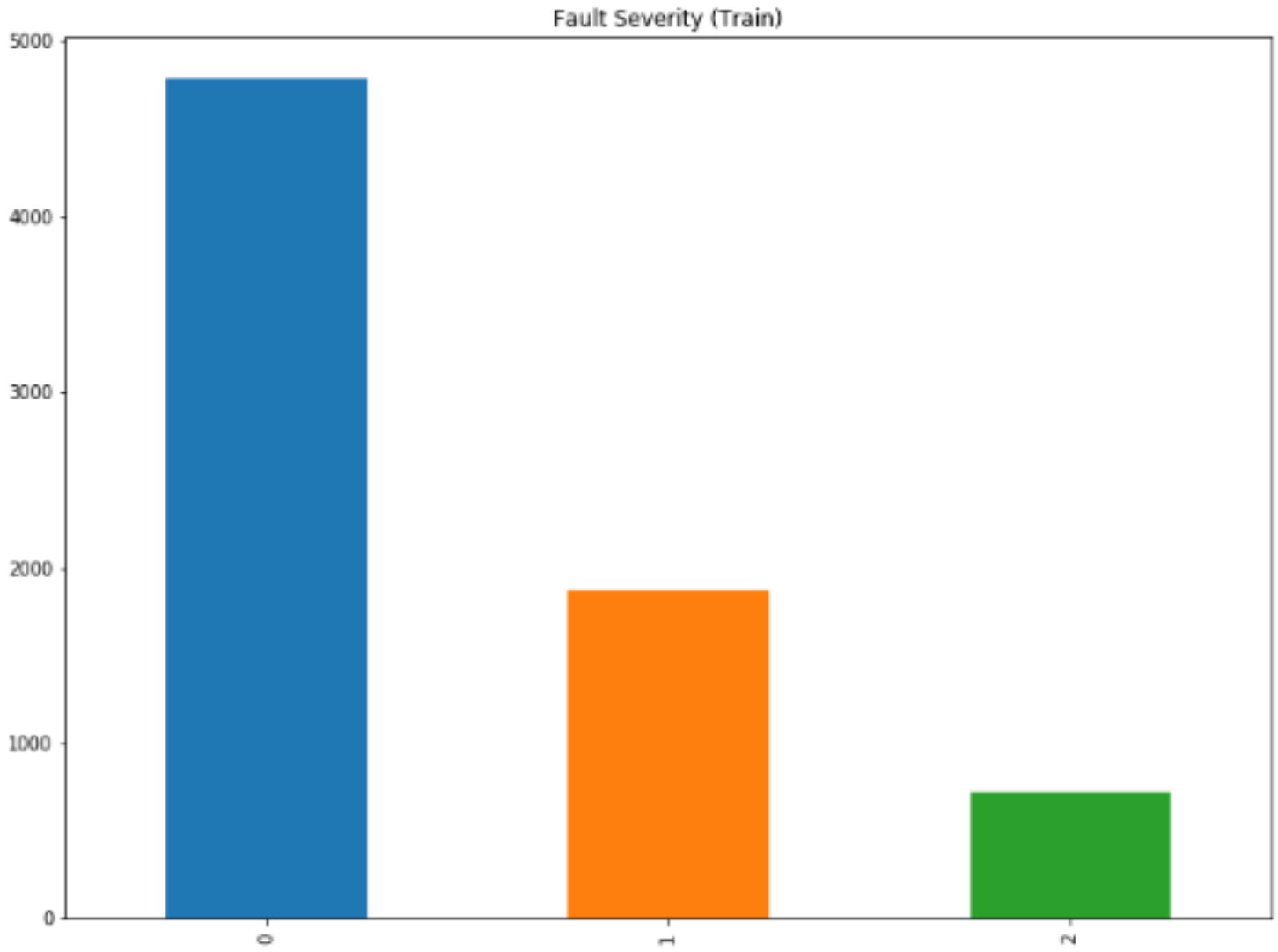
68
68, 172, 56, 193,

Data Dictionary

Data Dictionary is based on inferences drawn from data profiling and analysis

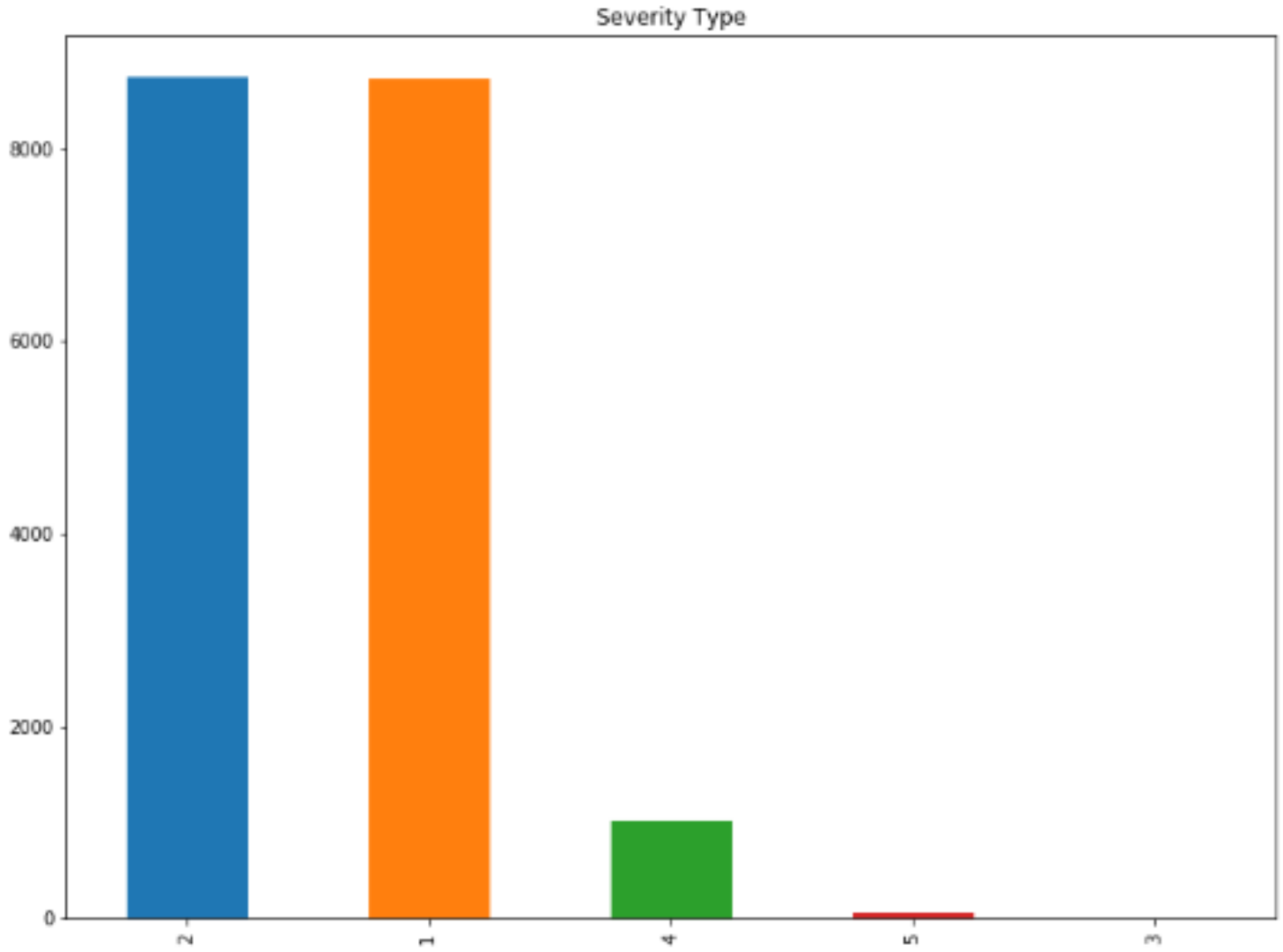
Entity	Log feature	
Description	Feature(s) associated with a log. Attribute volume gives the intensity of the feature. More than one log feature associated to a Log	
Attributes	Id	Unique Identifier for a log
	log_feature	Id for the log feature (no description)
	Multiple classes - values like 68, 172, 56, 193	
	volume	Intensity of the log feature
	(Values range from 1 to 1310)	
Comment	One data set has been provided	

Fault Severities



- 9.8% events lead to outage
- 65% events have no impact

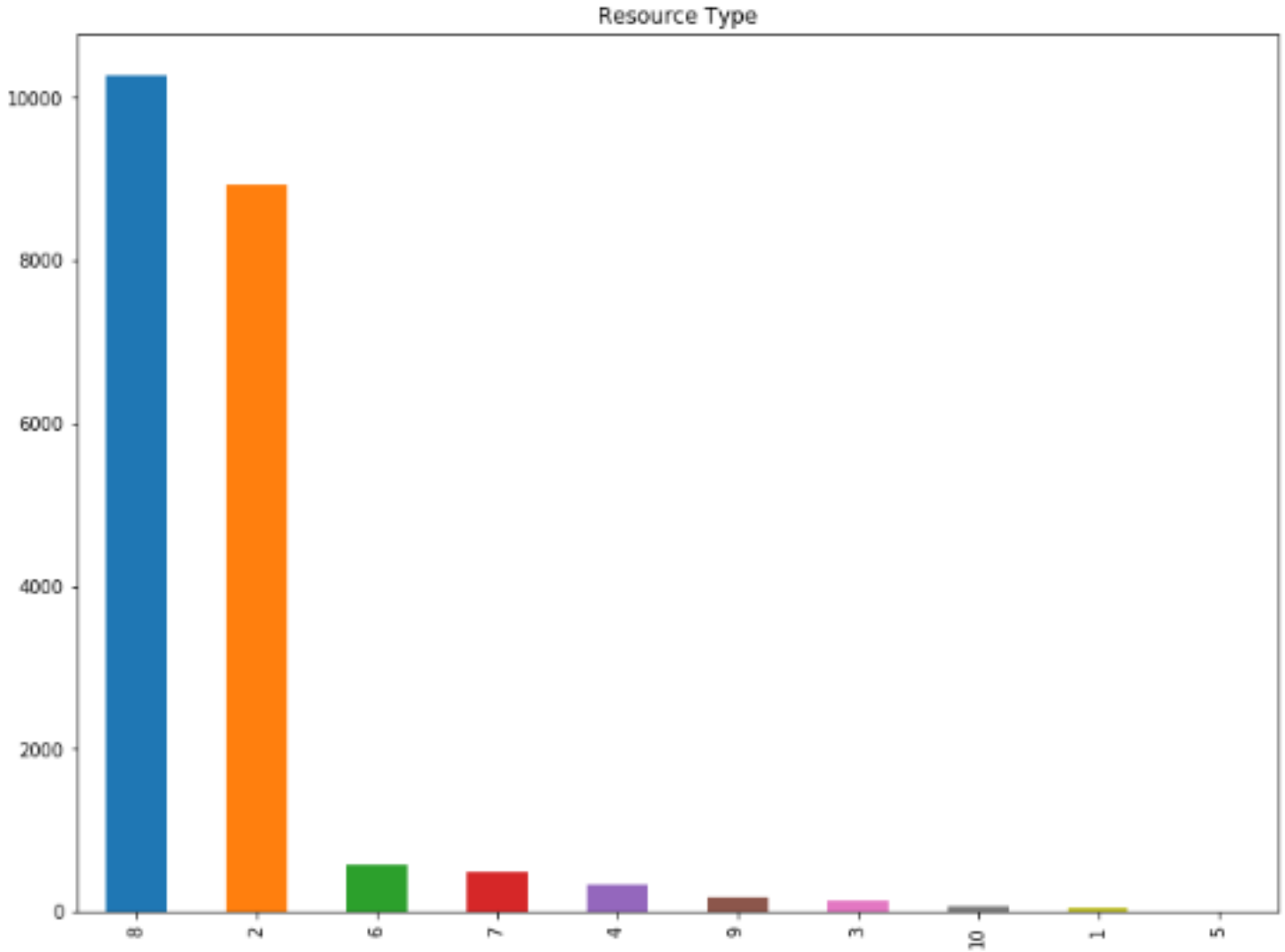
Severities



- Severity Type 1 & 2 have 85% of data
- Highly skewed to the right

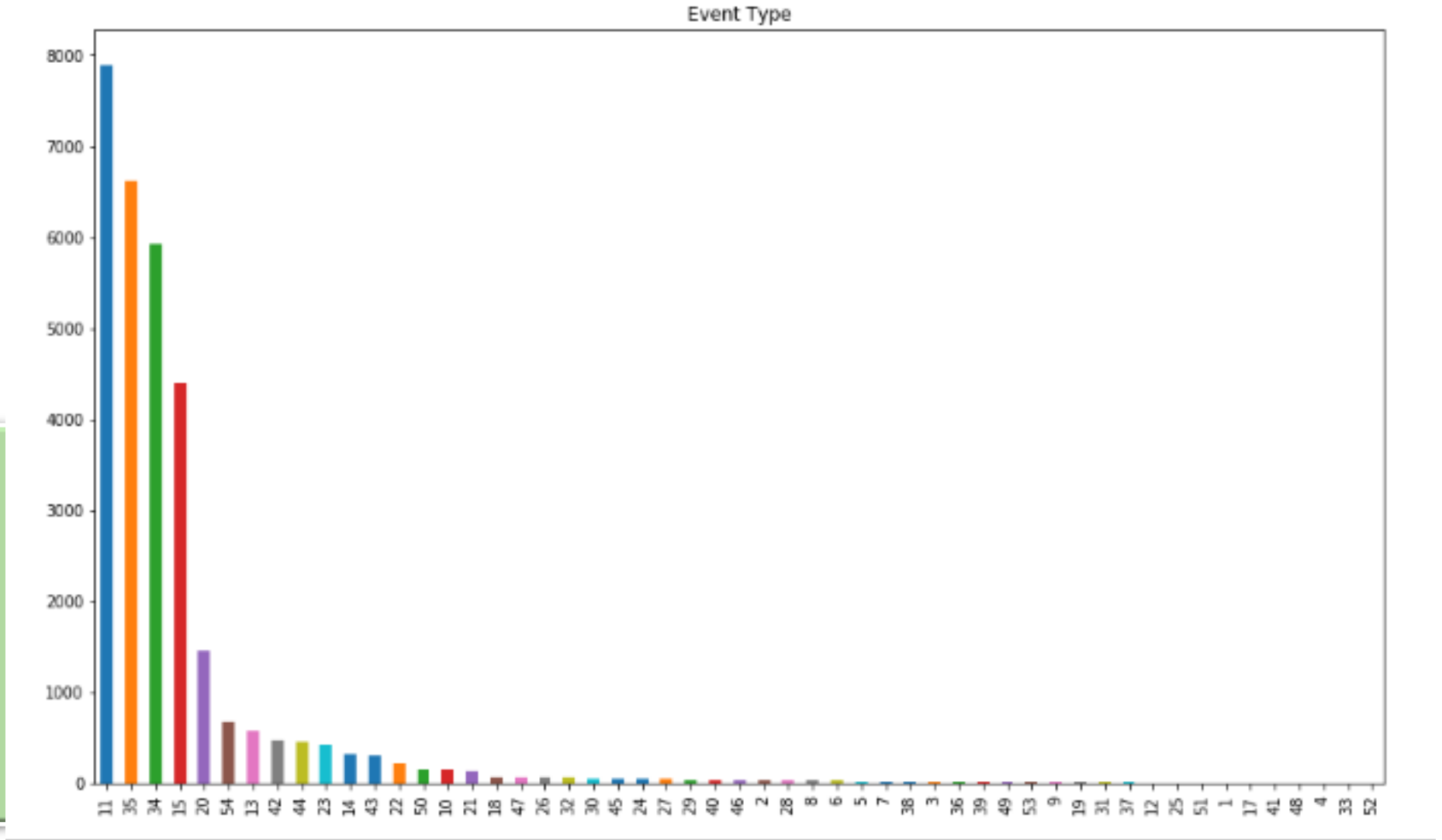
Resources

- Resource 8 has 49% of data
- Resource 2 has 42% of data
- Highly skewed to the right



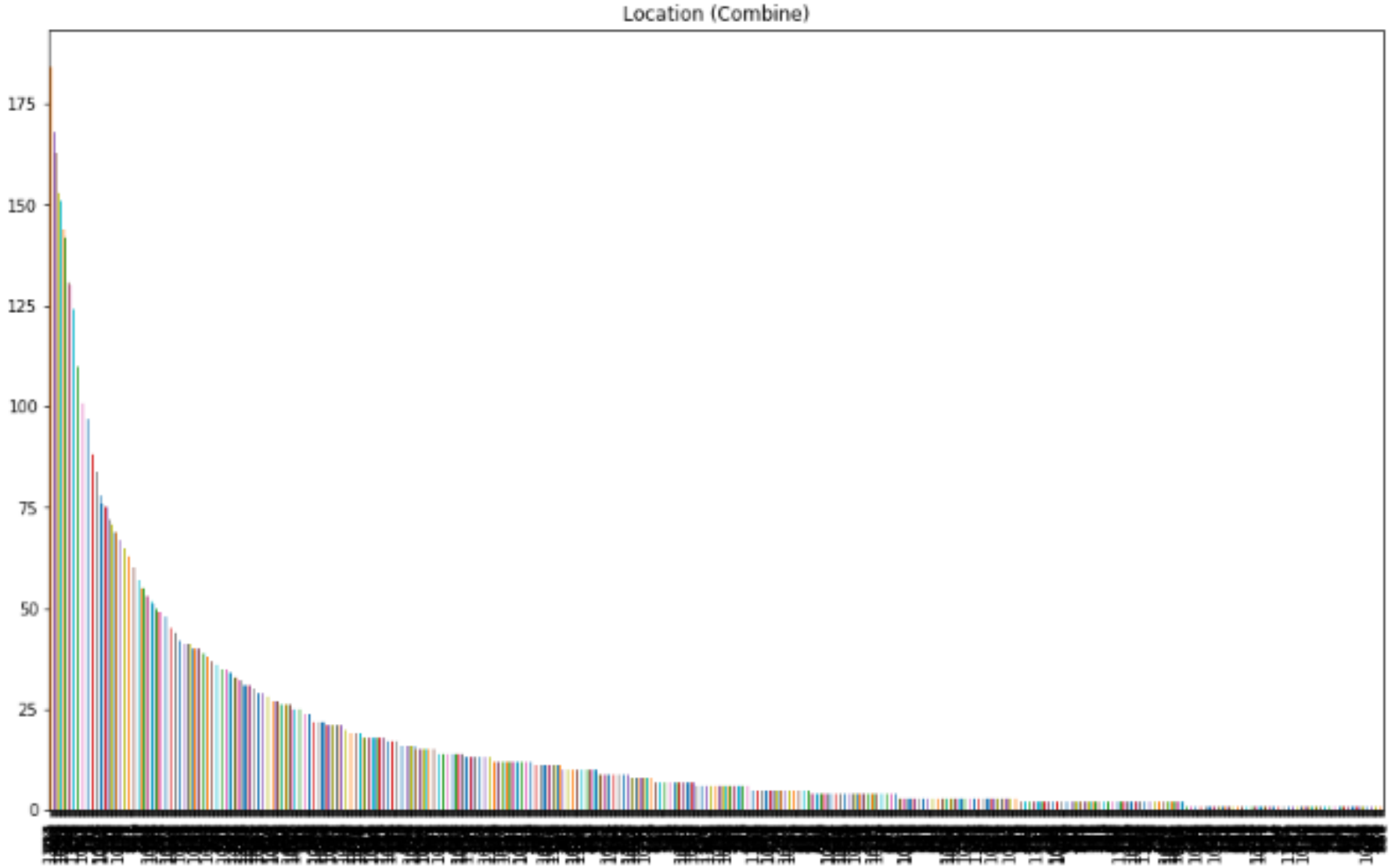
Events

- Top 3 events have 65% of data
- Top 5 events have 84% of data
- Highly skewed to the right



Locations

- Locations with frequency of 100 and more have 23% of data
- Highly skewed to the right but a larger spread

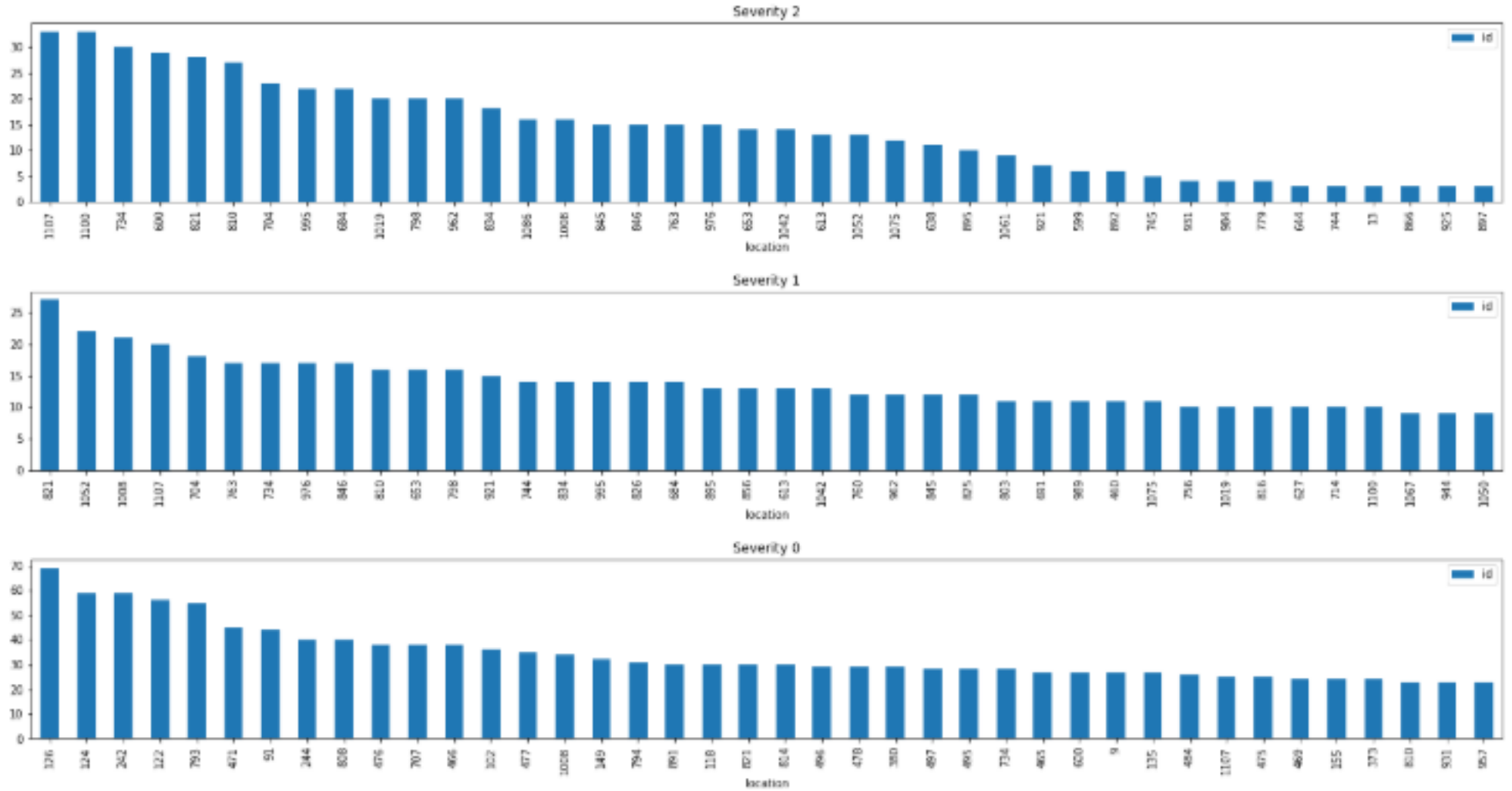


Locations by Fault severity

Total disruption

Momentary Glitch

No disruption



Data Engineering

- Almost all data is categorical
- The only exception is logfeature volume
- Distribution is biased for a few classes, data normalized by binning classes
- One hot encoding through the use of get_dummies method
- Logfeatures transposed to preserve volume

Resource Type

- Class 8 and class 2 retained
- Other classes binned into ‘OTH’
- Use *pd.get_dummies* to binarise data
- Group by ID to ensure one row for each ID

Event Type

- Classes 11,35,34,15,20 retained
- Other classes binned into ‘OTH’
- Use *pd.get_dummies* to binarise data
- Group by ID to ensure one row for each ID

Location

- Class retained if frequency >= 100
- Other classes binned into ‘OTH’
- Use *pd.get_dummies* to binarise data

Log feature

- Transposed log features to create 366 columns
- Map volume to each column

Severity Type

- Use *pd.get_dummies* to binarise data

Merge Data

- Dataframes merged with Train and Test dataframes
- Two dataframes created
  - All dataframes merged
  - Logfeatures excluded

Model Selection and Optimisation

Estimators

Estimators selected for evaluation

- sklearn.ensemble.RandomForestClassifier
- sklearn.ensemble.AdaBoostClassifier
- sklearn.ensemble.GradientBoostClassifier
- sklearn.neighbors.KNeighborsClassifier
- sklearn.tree.DecisionTreeClassifier

Hyperparameters

Steps for finalization of hyperparameters

- Instantiate model for each estimator
- Get full list of hyperparameters for each estimator using *get_params()* method
- Determine parameters influencing model accuracy using *sklearn.model_selection.RandomizedSearchCV*

Scenarios

Scenarios used for comparing model accuracy

- Full training dataset - Single run – Default parameters
- Training split 70% train – 30% validate using *sklearn.model_selection.train_test_split*
  - Run10 iterations using *sklearn.model_selection.cross_val_score* for
    - Model instantiated with default parameters
    - Model instantiated with hyperparameters

Model Accuracy Results

Estimator	All features				Features except logfeature		
	Single Run	Default params 10 iters	Optimum params 10 iters		Single Run	Default params 10 iters	Optimum params 10 iters
Random Forest Classifer	71.97%	70.30%	70.70%		59.41%	57.90%	61.10%
KNeigbors Classifer	59.91%	64.00%	64.00%		59.37%	59.80%	64.40%
DecisionTree Classifer	69.35%	68.90%	68.90%		59.10%	56.20%	56.70%
AdaBoost Classifer	72.42%	71.30%	71.30%		66.23%	64.90%	64.90%
GradientBoost Classifer	76.30%	73.70%	73.90%		66.41%	66.20%	64.60%

Model Accuracy Inference

- 5 classification estimators used
- 20 scenarios used for comparing model accuracy
- “logfeature” is a significant feature as it affects accuracy
- GradientBoost Classifier is the best estimator overall
  - 76.30% accuracy for single run
  - 73.90% accuracy for 10 iterations with hyperparameters
  - 73.70% accuracy for 10 iterations with default parameters
- **Recommend GradientBoost Classifier**

Top 20 Features

Random Forest			Decision Tree			AdaBoost			GradientBoost	
log_feature203	0.117975		log_feature203	0.196701		log_feature203	0.1		log_feature203	0.124773
log_feature82	0.083862		severity_type_1	0.058009		log_feature170	0.06		log_feature170	0.034011
log_feature170	0.03621		log_feature82	0.050477		resource_type_RT8	0.06		log_feature202	0.03206
log_feature54	0.033386		log_feature170	0.044608		event_type_OTH	0.04		log_feature209	0.024589
log_feature232	0.027159		log_feature54	0.02519		log_feature202	0.04		log_feature232	0.024231
log_feature312	0.022933		log_feature312	0.024538		location_995	0.02		log_feature312	0.023538
event_type_OTH	0.022118		log_feature80	0.022101		location_OTH	0.02		log_feature73	0.023496
log_feature80	0.021604		log_feature68	0.019273		event_type_ET11	0.02		log_feature82	0.018607
log_feature68	0.020152		log_feature232	0.017752		event_type_ET34	0.02		log_feature171	0.018412
log_feature71	0.018804		resource_type_OTH	0.015536		event_type_ET35	0.02		log_feature155	0.016335
location_OTH	0.016184		event_type_OTH	0.014884		severity_type_1	0.02		log_feature179	0.016276
event_type_ET15	0.016145		log_feature73	0.014877		log_feature193	0.02		severity_type_1	0.014595
event_type_ET34	0.015661		log_feature71	0.013795		log_feature195	0.02		log_feature134	0.01443
severity_type_1	0.015401		log_feature171	0.012685		log_feature196	0.02		log_feature315	0.014178
log_feature313	0.014771		log_feature315	0.012159		log_feature205	0.02		log_feature70	0.014025
log_feature201	0.014159		log_feature193	0.011945		log_feature140	0.02		log_feature368	0.013464
log_feature193	0.013333		log_feature201	0.011271		log_feature209	0.02		log_feature227	0.012689
severity_type_2	0.012206		log_feature291	0.011234		log_feature212	0.02		log_feature314	0.012604
log_feature73	0.011528		event_type_ET11	0.009957		log_feature319	0.02		log_feature54	0.012336
resource_type_RT8	0.011016		event_type_ET15	0.00971		log_feature295	0.02		event_type_OTH	0.012102