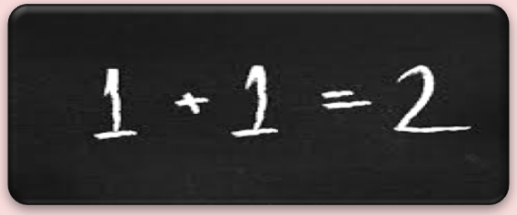
Everything should be made as simple as possible, but not simpler

Albert Einstein

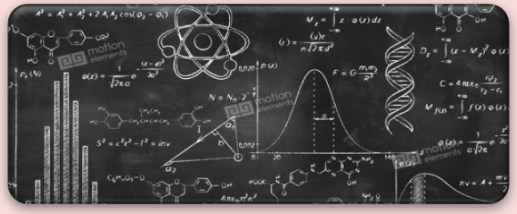
MODEL VS FEATURE COMPLEXITY

Xuan Qi

Project design



$$9 - 3 \div \frac{1}{3} + 1 =$$



Scenario 1

- Simple features
- Simple models

Scenario 2

- Complex features
- Simple models

Scenario 3

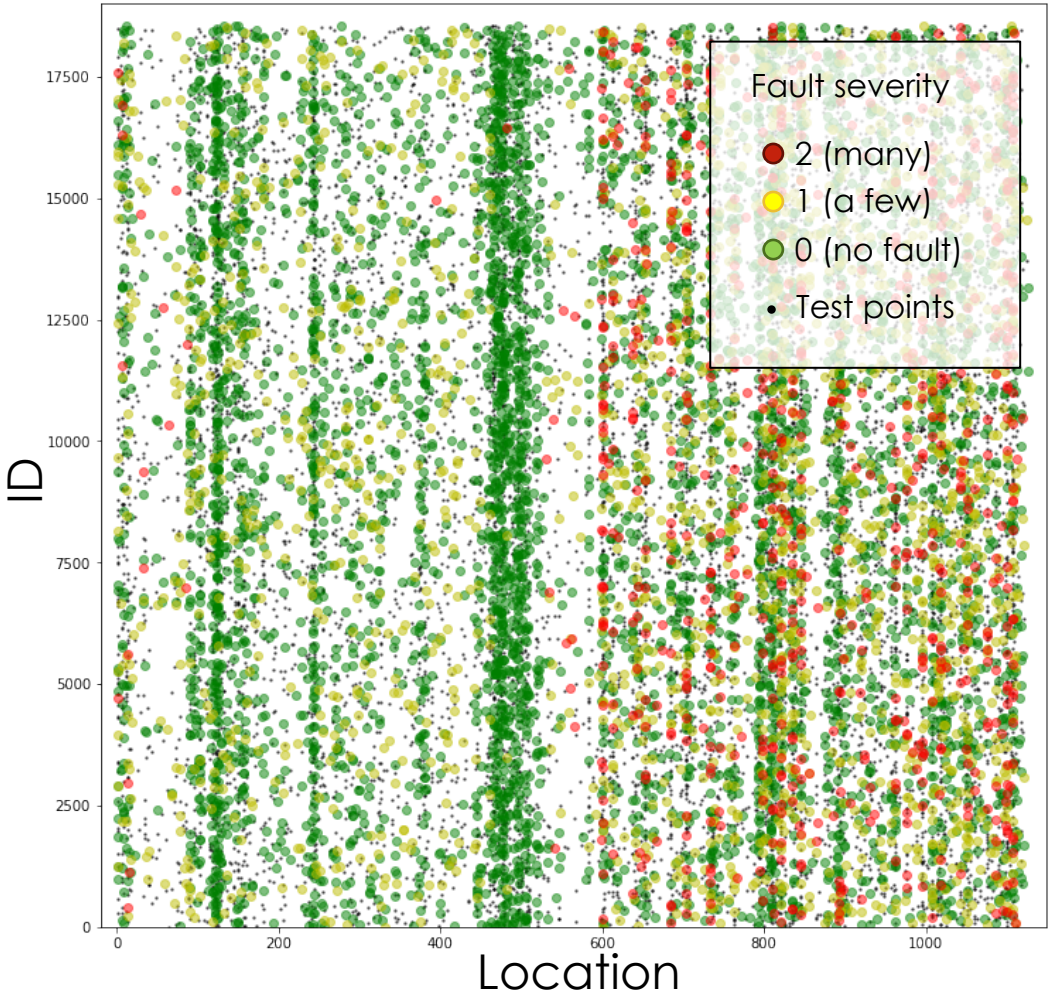
- Complex features
- Complex models

DATA

- **Telstra network disruptions**

- Use a dataset of features from service logs
- Predict if a disruption is a momentary glitch or a total interruption of connectivity

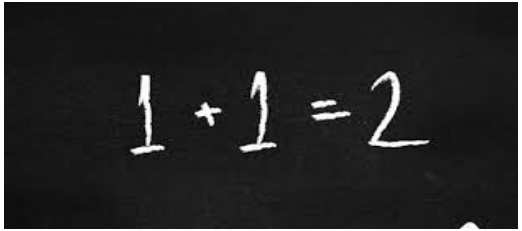
https://www.kaggle.com/c/telstra-recruiting-network



TELSTRA DATA

- Target: Fault severity
- Features:
  - Event type (20)
  - Severity type (5)
  - Resource type (5)
  - Log feature (147)
  - Log volume
- Evaluate:
  - multi-class logarithmic loss

$$logloss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}\log(p_{ij})$$



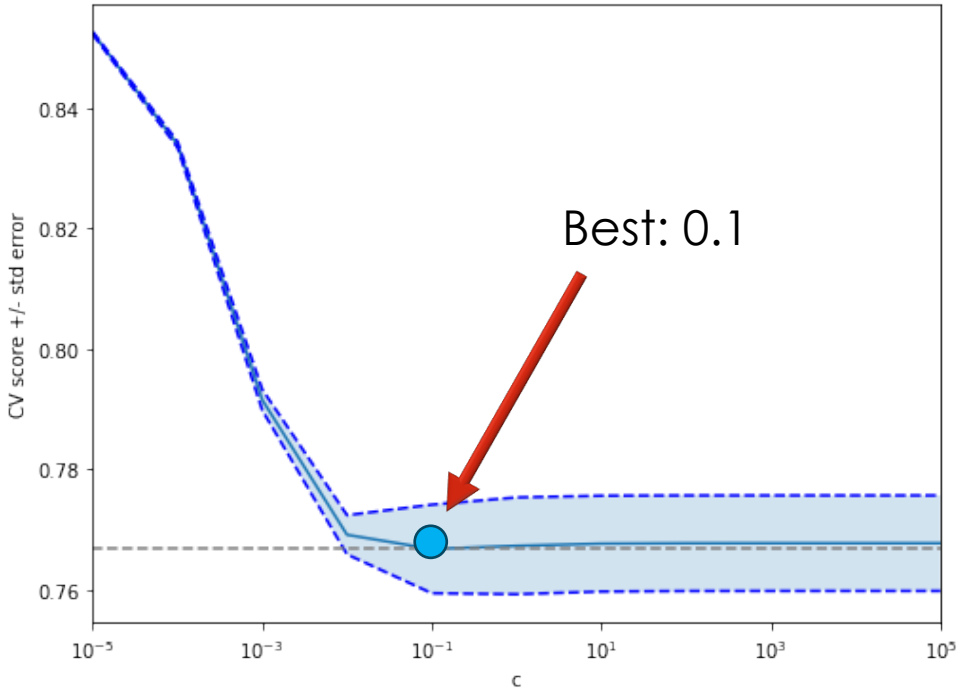
- Simple feature:
  - Features (10):

    location, severity type, event count, resource count, log volume count, log volume min, log volume mean, log volume max, log volume std, log volume sum

- Simple model:
  - Logistic regression
  - Support vector machine (SVM)
  - Random forest

Logistic regression

Tuning parameters: C
Log loss: 0.769



SVM

Tuning parameters: C, gamma, kernel
Log loss: 0.708

	0.001	0.01	0.1	0	1	10	100
0.01	0.804	0.774	0.769	0.855	0.768	0.794	0.79
0.1	0.806	0.784	0.764	0.855	0.766	0.794	0.789
1	0.795	0.769	0.733	0.855	0.726	0.748	0.782
10	0.789	0.747	0.713	0.855	0.72	0.76	0.795
100	0.778	0.725	0.708	0.855	0.733	0.784	0.811

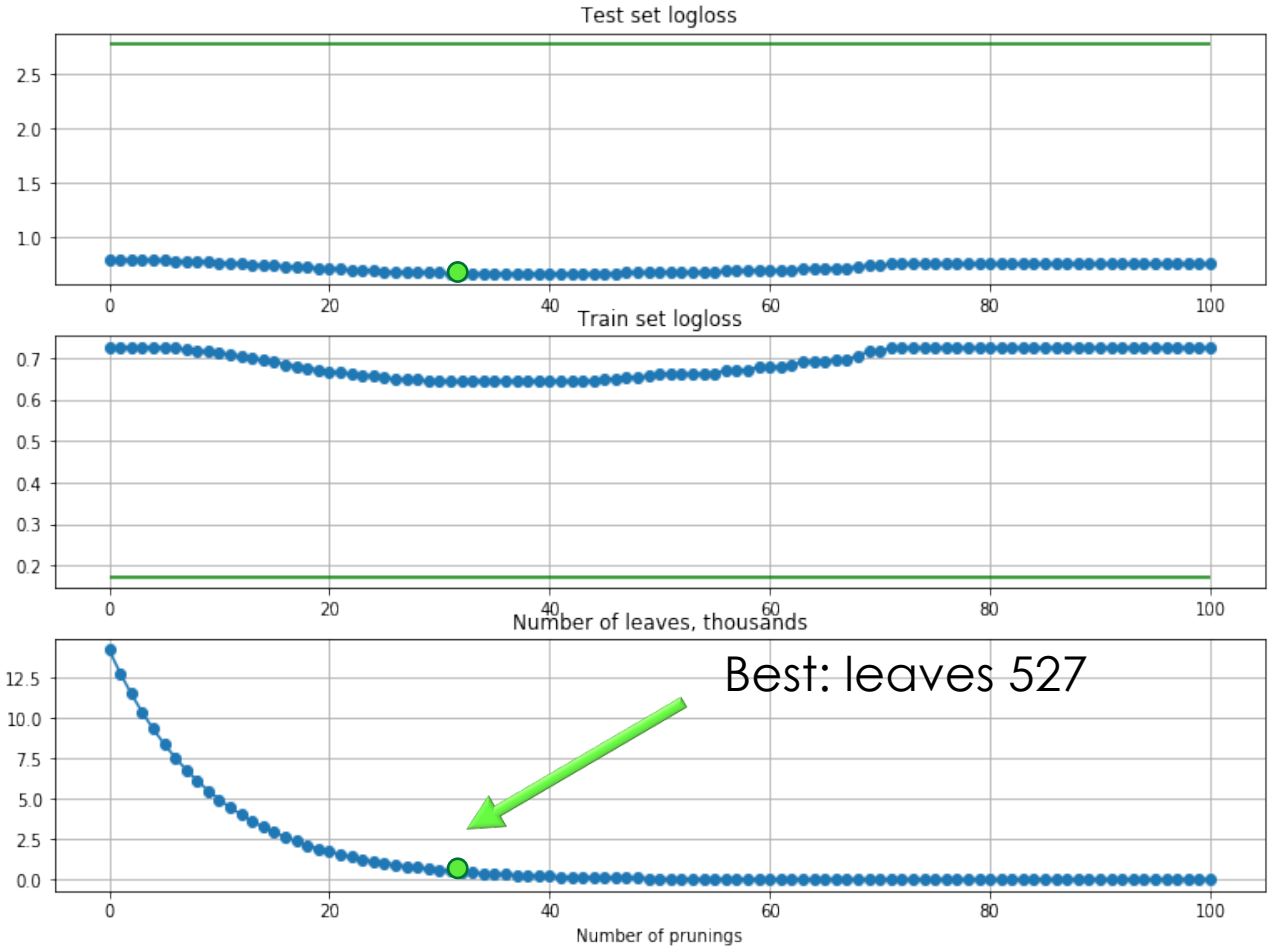
Best: c=100, gamma = 0.1, kernel: rbf

Random forest

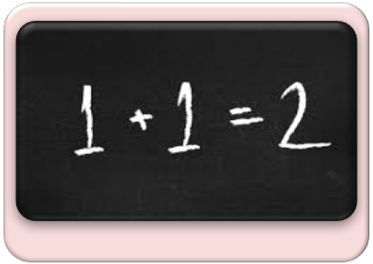
	Training log loss		Testing log loss
No pruning	0.1735	<<	2.7796
With pruning	0.6440		0.6684

Over training !

Random forest

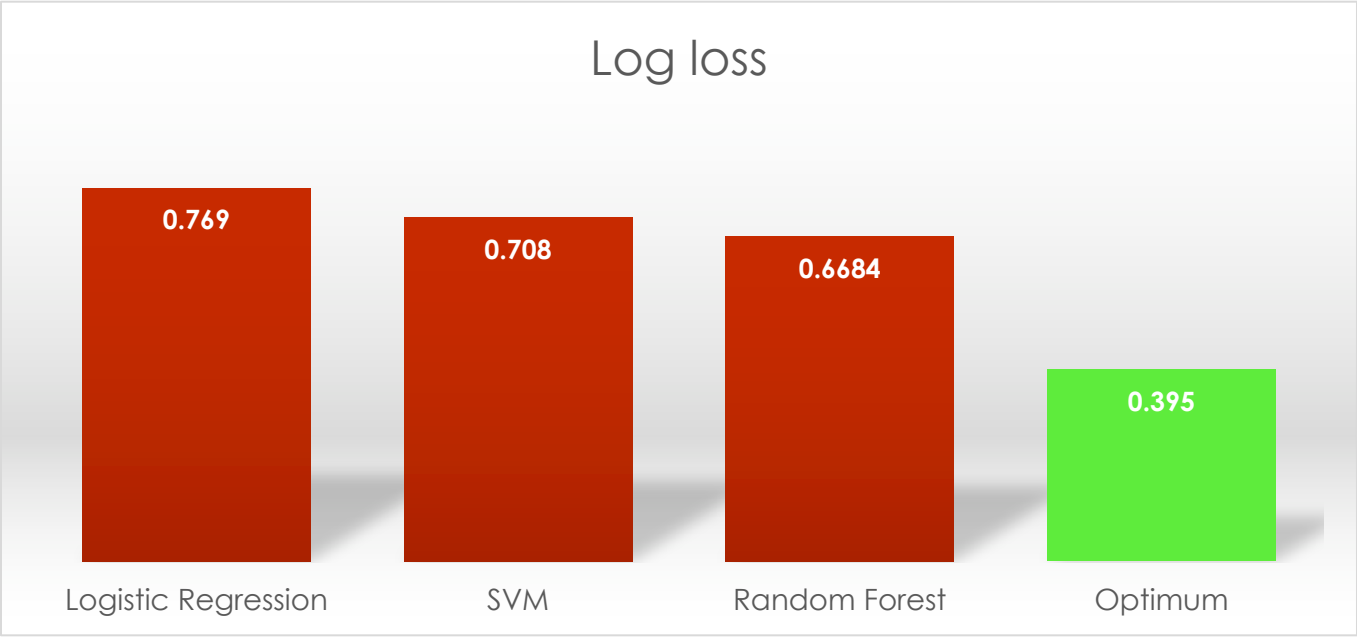


SCENARIO 1

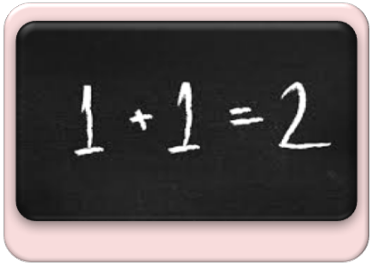


Scenario 1

- Simple data
- Simple models

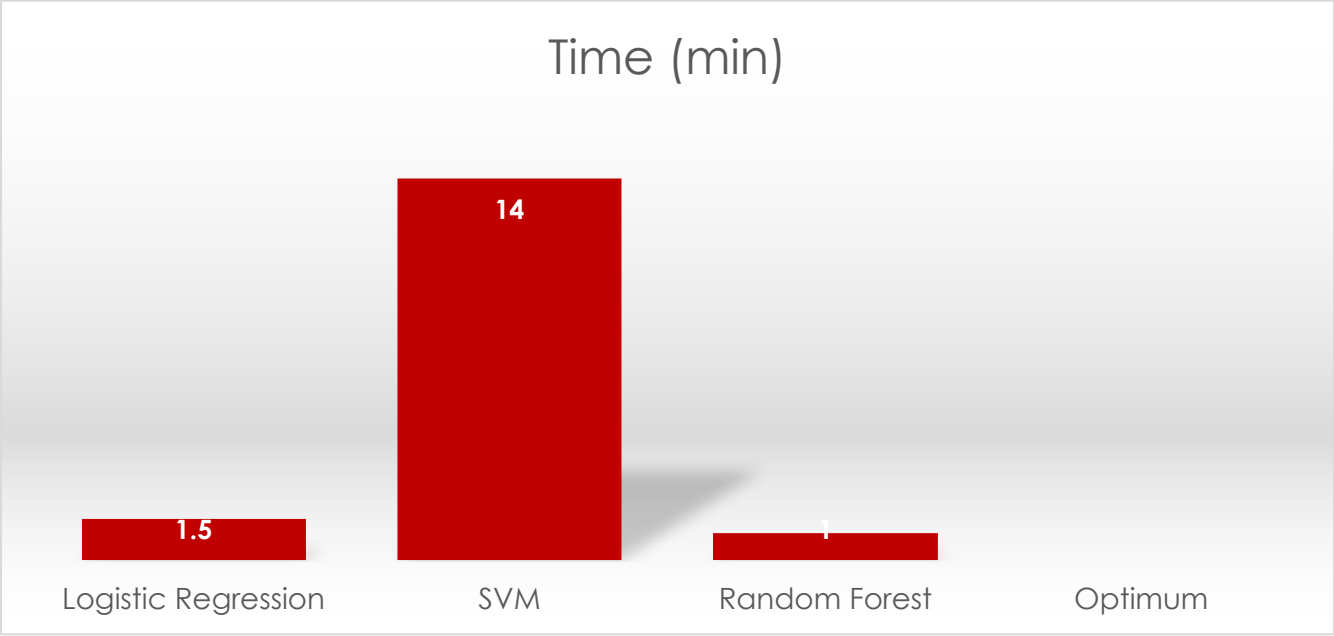


SCENARIO 1



Scenario 1

- Simple data
- Simple models



Can You Solve This?

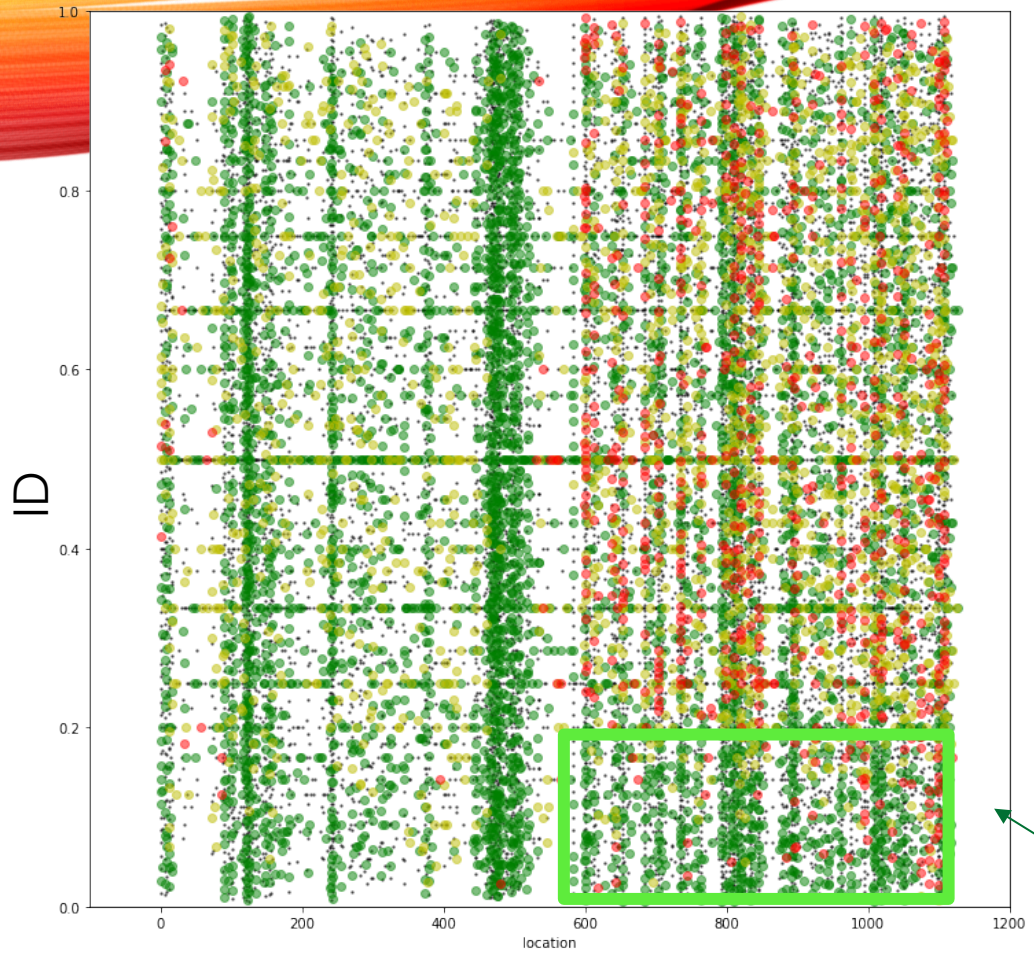
$$9 - 3 \div \frac{1}{3} + 1 =$$

- Complex features:
  - Features (172):

    Magic time feature, location, severity type, event count, resource count, log volume count, log volume min, log volume mean, log volume max, log volume std, log volume sum, log features, the probabilities of fault severities at every location etc.

- Simple model:
  - Logistic regression
  - Support vector machine (SVM)
  - Random forest

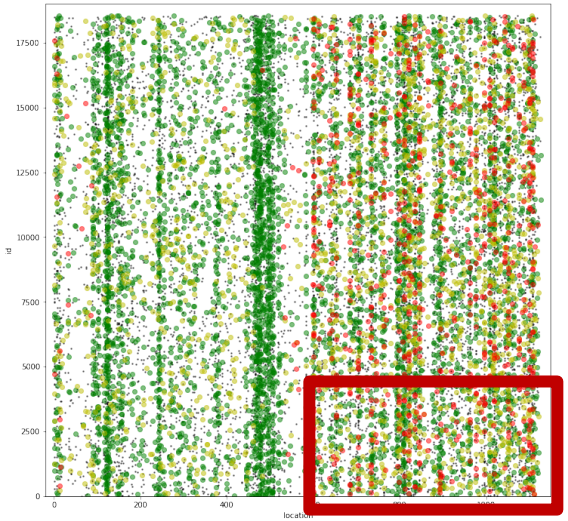
Magic time feature



Location

Fault severity

- 2 (many)
- 1 (a few)
- 0 (no fault)
- Test points



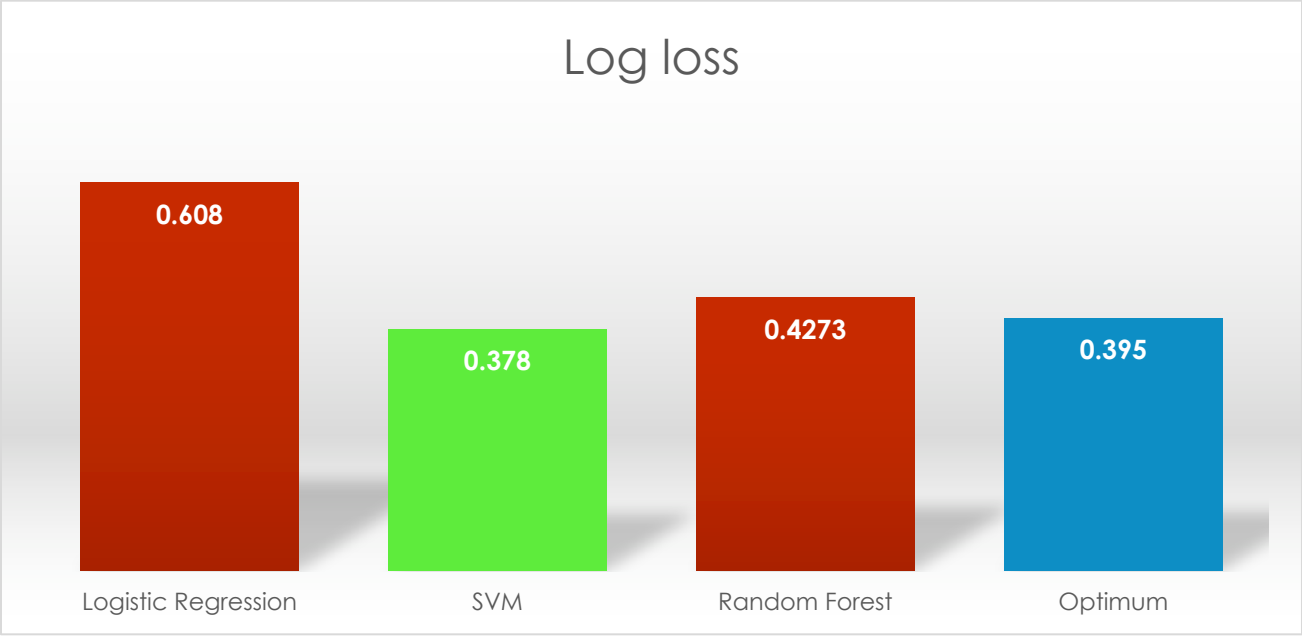
No fault

SCENARIO 2

$$9 - 3 \div \frac{1}{3} + 1 =$$

Scenario 2

- Complex data
- Simple models

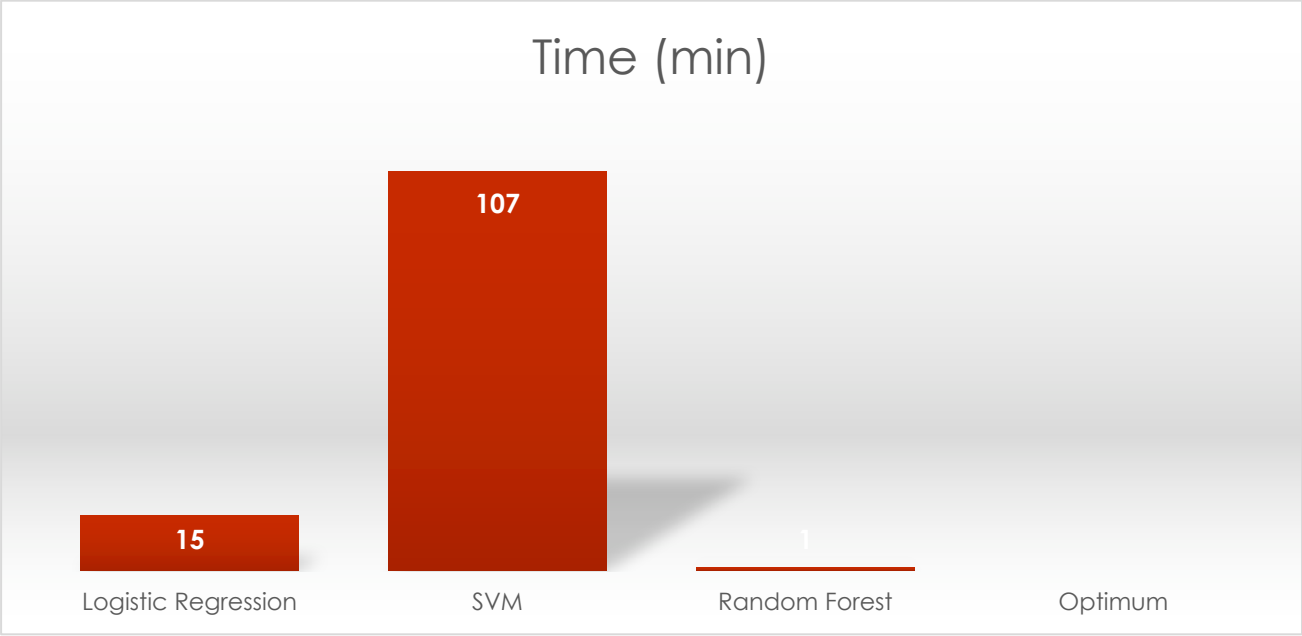


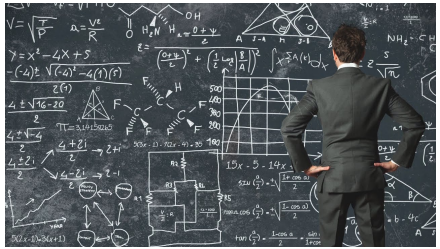
SCENARIO 2

$$9 - 3 \div \frac{1}{3} + 1 =$$

Scenario 2

- Complex data
- Simple models



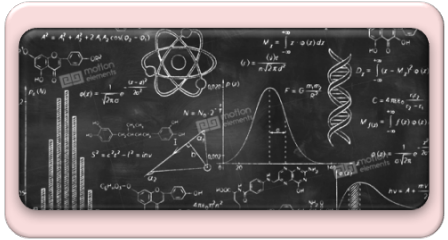


- Complex features:
  - Features (172):

    Magic time feature, location, severity type, event count, resource count, log volume count, log volume min, log volume mean, log volume max, log volume std, log volume sum, log features, the probabilities of fault severities at every location etc.

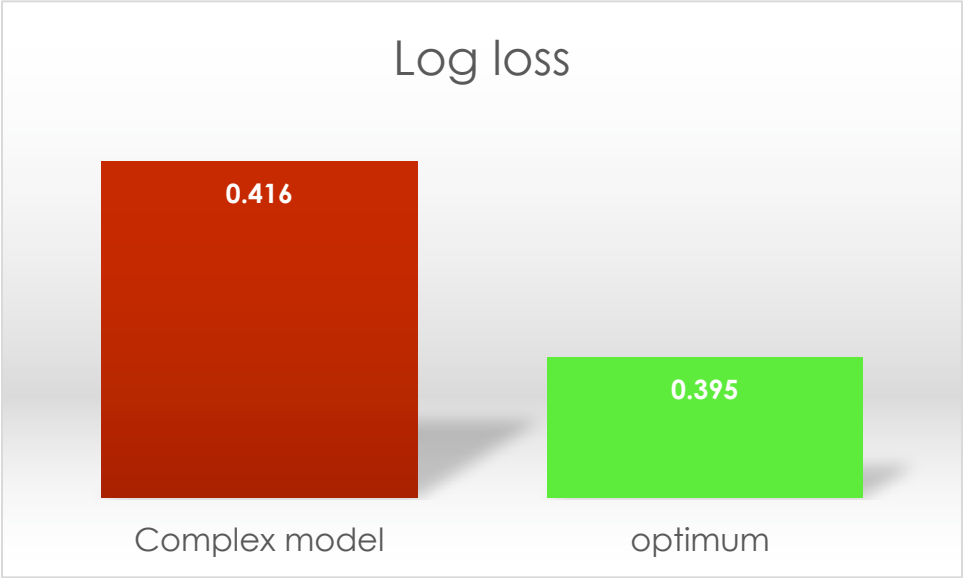
- Complex model:
  - Neural network
  - XGBoost models

SCENARIO 3

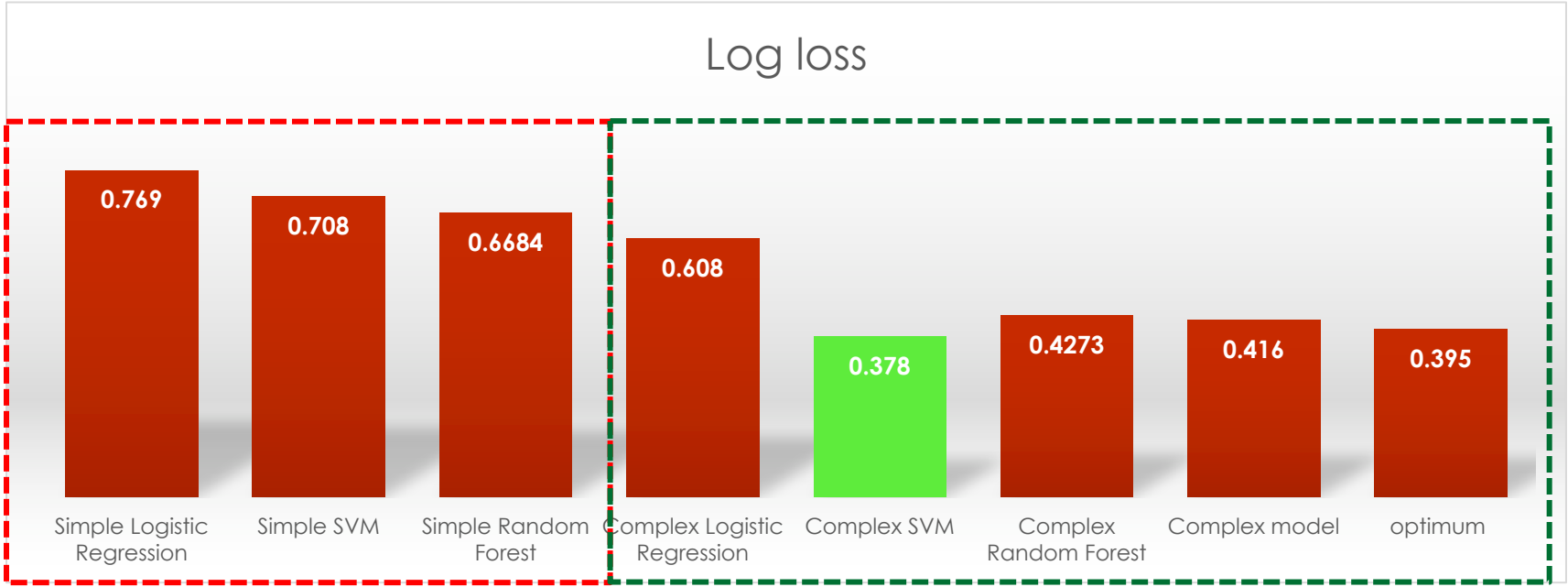


Scenario 3

- Complex features
- Complex models

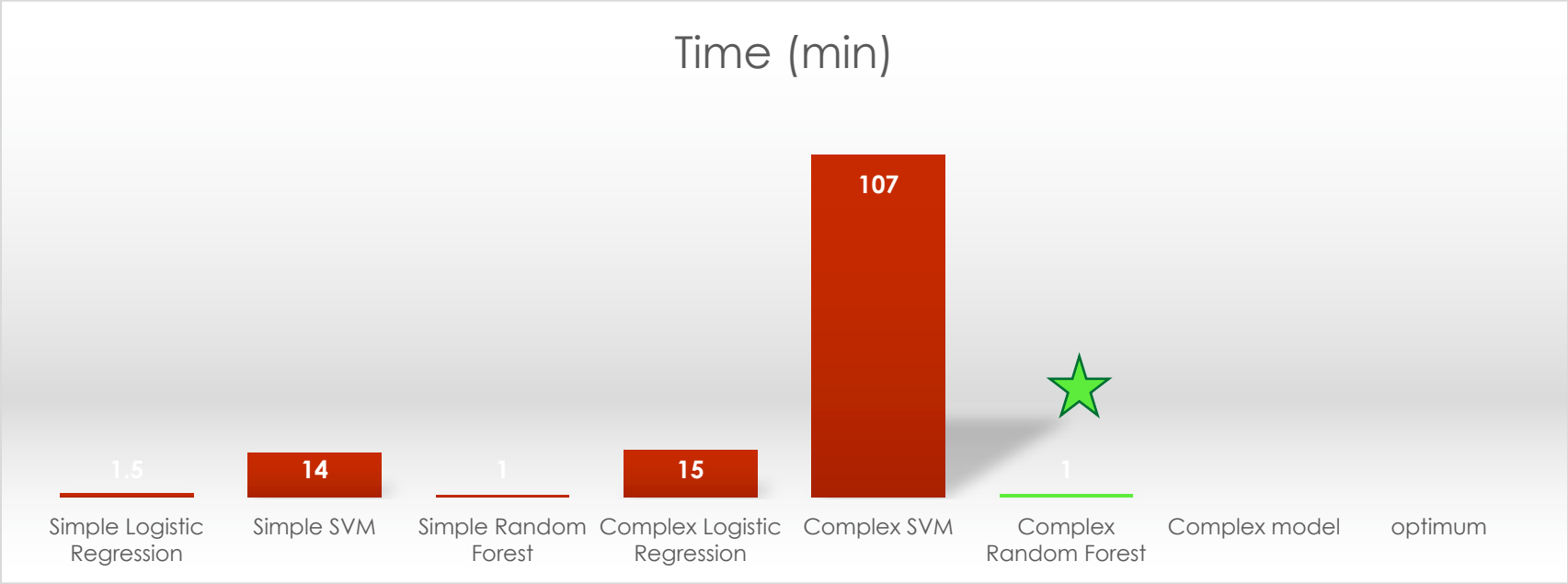


CONCLUSION



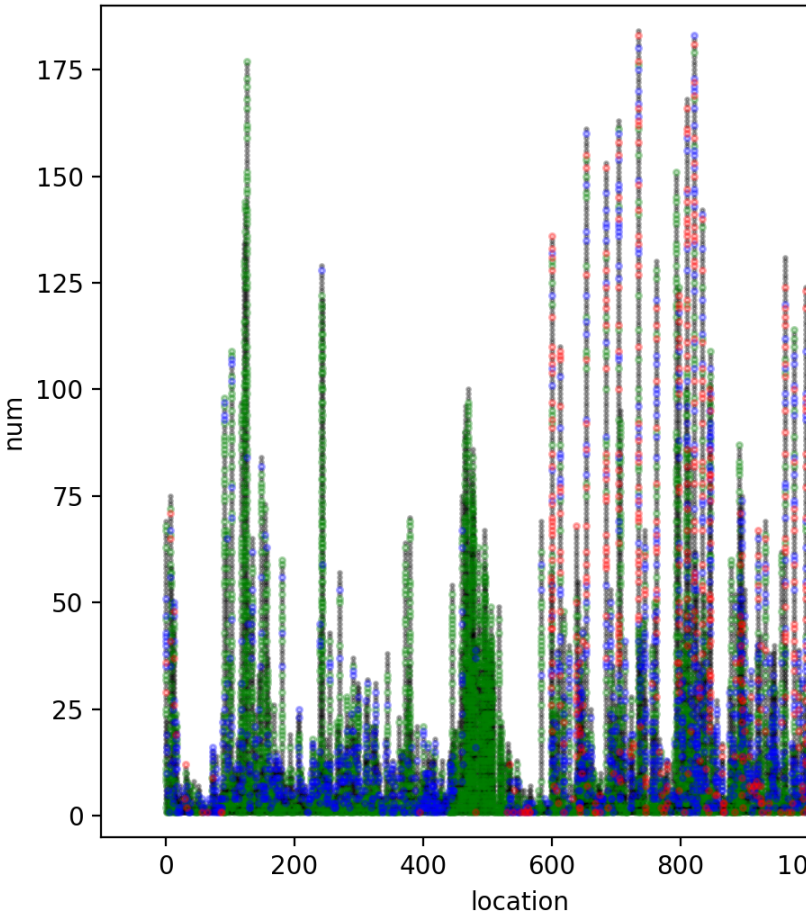
Simple features **Complex features**

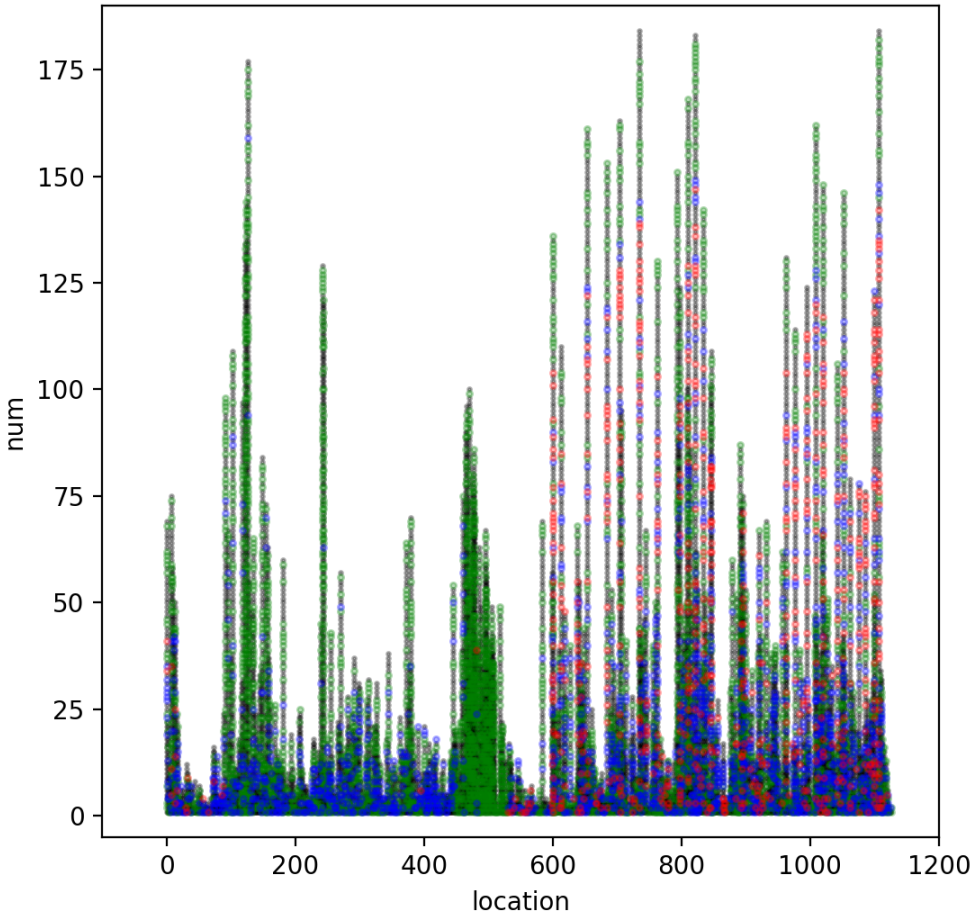
CONCLUSION


Time (min)
1.5
14
1
15
107
1
Simple Logistic Regression
Simple SVM
Simple Random Forest
Complex Logistic Regression
Complex SVM
Complex Random Forest
Complex model
optimum

THANK YOU

Questions?


175
150
125
100
75
50
25
0
num
0
200
400
600
800
10
location


175
150
125
100
75
50
25
0
num
0
200
400
600
800
1000
1200
location