Seismic Data Analysis

Presented By - The Geeks

1.Prabhat Kumar Sahu
2.Batakrishna Sahu
3.Suman Mishra
4.Nikita Jit

Problem Statement

Data Analysis and Development of Environmental Information System for Mining Areas in Odisha.

What is Data ?

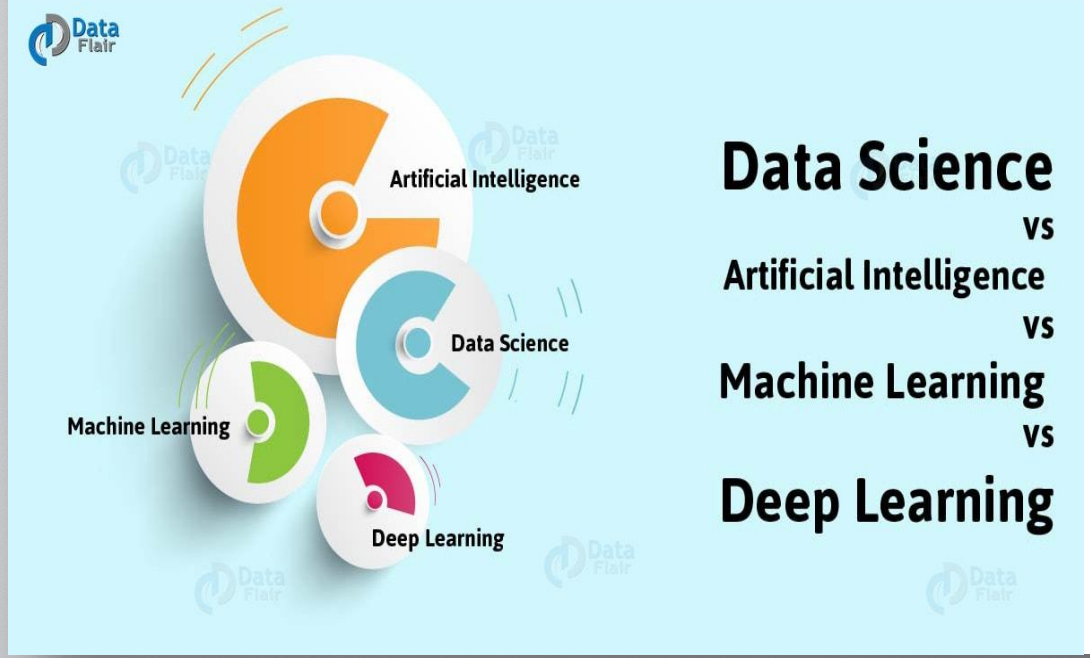
In general, data is any set of characters that has been gathered and translated for some purpose, usually analysis. It can be any character, including text and numbers, pictures, sound, or video. If data is not put into context, it doesn't do anything to a human or computer.

Why Data Is Important ?

- Used to help decision-making
- Predictive analytics
- Can reveal patterns and trends and other crucial facts
- Business Intelligence

You name any thing, the word Data is associated with it.

The Buzz Words



Data analysis :

Data analysis is a process of inspecting, cleansing, transforming, and modeling data with the goal of discovering useful information, informing conclusions, and supporting decision-making. Data analysis has multiple facets and approaches, encompassing diverse techniques under a variety of names, while being used in different business, science, and social science domains.

How big organization use Machine Learning & Data Analysis ?

- Recommendation
  - Ecommerce (Flipkart,Amazon,Myntra..etc)
  - Netflix (Movie, Tv Series)
  - Spotify (Music)
  - Youtube (Relevant Videos)
  - Facebook news feed
- Fraud and Risk Detection
- Internet/Web Search
- Digital Advertisement
- Customer Interactions
- Delivery logistics
- Weather Prediction

More Use Cases

- Image Recognition
- Speech Recognition
- Gaming
- Smart Chatbots
- Robotics
- Self Driving Cars
- Healthcare (Medical Imaging ..etc)
- Intelligent transport systems
- Congestion management by predicting traffic conditions

Agenda :

Mines collect a lot of data but they don’t analyse and get the proper insights out of it.

With the information about the possibility of hazardous situation occurrence, an appropriate supervision service can reduce a risk of rockburst .

Solution :

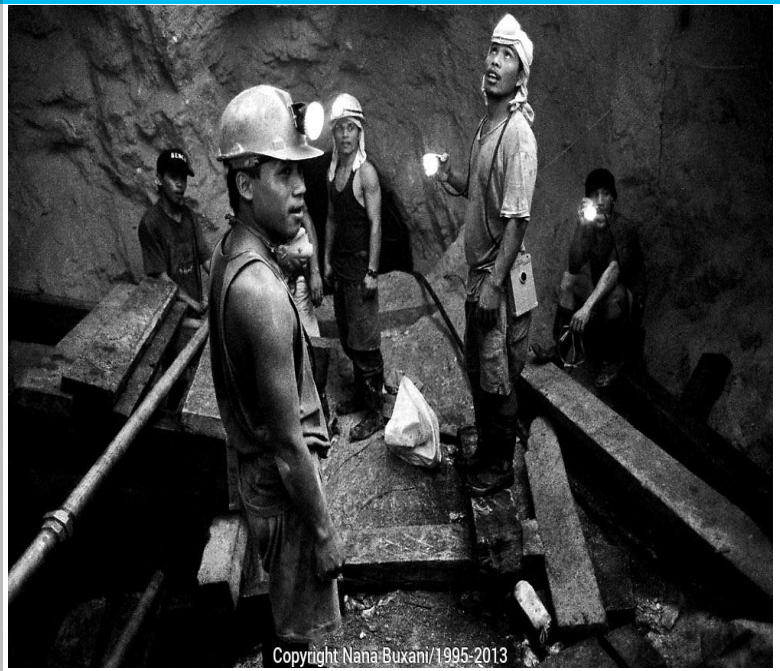
Our analysis attempts to use logistic regression techniques to predict whether a seismic 'bump' is predictive of a notable seismic hazard. We attempt to characterize our prediction accuracy and compare the results against the state of the art results from other statistical and machine learning techniques, that are included within the data set.

What Is Seismic Data ?

- Mining activity has long been associated with mining hazards, such as fires, floods, and toxic contaminants .
- Among these hazards, seismic hazards are the hardest to detect and predict.Minimizing loss from seismic hazards requires advanced data collection and analysis.
- In recent years, more and more cutting-edge seismic and seismo-acoustic monitoring systems have come about. Still, the disproportionate number of low-energy versus high-energy seismic phenomena renders traditional analysis methods insufficient in making accurate predictions.

How People Work At Coal Mines?

People Working Under Mines



Seismic Data Analysis : Introduction

- The dangers associated with coal mining are myriad; black lung, flammable gas pockets, rock-bursts, and tunnel collapses are all very real dangers that mining companies must consider when attempting to provide safe working conditions for miners.
- One class of mining hazard, commonly called **'seismic hazards'**, are notoriously difficult to protect against and even more difficult to predict with certainty. Therefore, predicting these hazards has become a well-known problem for machine learning and predictive analytics.

Attribute Information

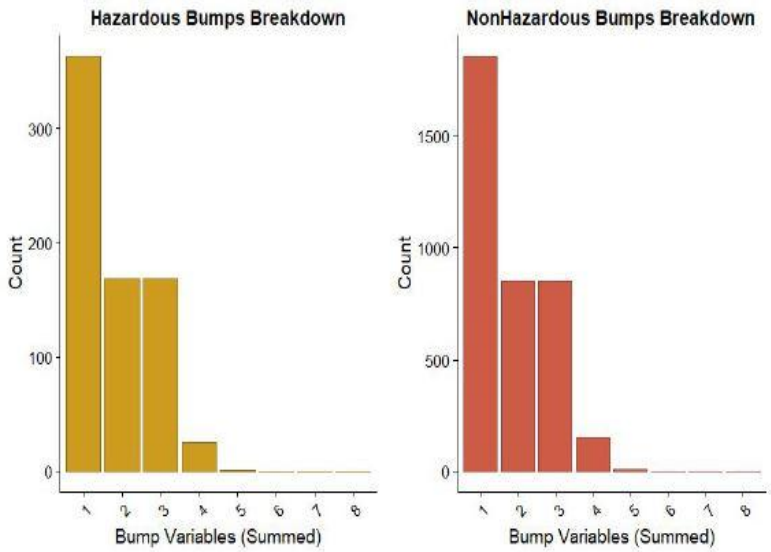
1. seismic: result of shift seismic hazard assessment in the mine working obtained by the seismic method (a - lack of hazard, b - low hazard, c - high hazard, d - danger state);	2. seismoacoustic: result of shift seismic hazard assessment in the mine working obtained by the seismoacoustic method;
3. shift: information about type of a shift (W - coal-getting, N -preparation shift);	4. genergy: seismic energy recorded within previous shift by the most active geophone (GMax) out of geophones monitoring the longwall;
5. gpuls: a number of pulses recorded within previous shift by GMax;	6. gdenergy: a deviation of energy recorded within previous shift by GMax from average energy recorded during eight previous shifts;
7. gdpuls: a deviation of a number of pulses recorded within previous shift by GMax from average number of pulses recorded during eight previous shifts;	8. ghazard: result of shift seismic hazard assessment in the mine working obtained by the seismoacoustic method based on registration coming form GMax only;
9. nbumps: the number of seismic bumps recorded within previous shift;	10. nbumps2: the number of seismic bumps (in energy range [10^2,10^3)) registered within previous shift;
11. nbumps3:..7 the number of seismic bumps (in energy range [10^3..6,10^4..7)) registered within previous shift;	12. nbumps89: the number of seismic bumps (in energy range [10^8,10^10)) registered within previous shift;
13. energy: total energy of seismic bumps registered within previous shift;	14. maxenergy: the maximum energy of the seismic bumps registered within previous shift;
15. class: the decision attribute - '1' means that high energy seismic bump occurred in the next shift ('hazardous state'), '0' means that no high energy seismic bumps occurred in the next shift ('non-hazardous state').	

Dataset Description

- The data were taken from instruments in the Zabrze-Bielszowice coal mine, in Poland.
- There are 2,584 records, with only 170 class = 1 variables, so the data are significantly skewed towards non-hazardous training data.
- Essentially energy readings and bump counts during one work shift are used to predict a 'hazardous' bump during the next shift
- From the data description, a 'hazardous bump' is a seismic event with > 10,000 Joules, and a 'shift' is a period of 8 hours.
- For the sake of reference, a practical example of 10,000 Joules would be the approximate energy required to lift 10,000 tomatoes 1m above the ground.
- A class = 1 variable result signifies that a hazardous bump did, indeed, occur in the following shift to the measured data.

Main Effects

- The main effects for this study are considered to be the all numeric variables, plus ghazard, seismoacoustic and shift.
- The nbumps class of variables are left out for more advanced models
- Since the resonance and frequency ranges could have a multitude of confounding variables that we, without significant mining expertise, would miss.
- To test that nbumps isn't necessarily the largest effect, we looked at a side-by-side histogram of nbump records for each of the two output classes:
- The pattern of frequency distributions appears to be consistent, regardless of class. Therefore, we will not be making 'nbumps' one of our main effects variables.



Logistic Regression Assumptions

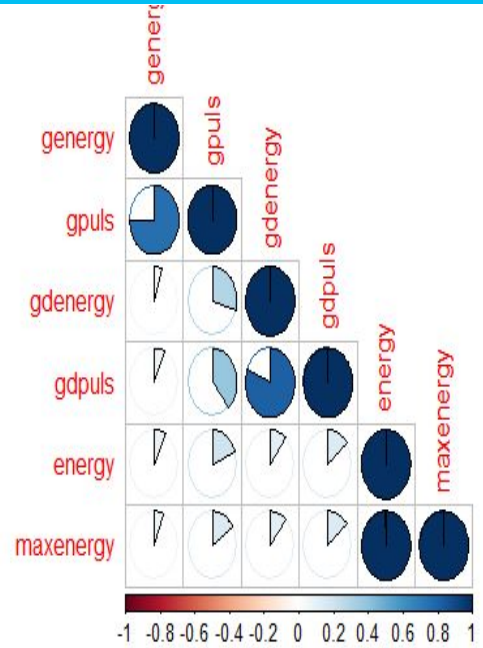
Output Variable :

Logistic regression requires a categorical output variable. In this case, our output variable (class) is a binary categorical variable.

Independence of Observations:

We are making the assumption that each measurement is an independent measurement, taken at different times, from the same mine. This is based on the data set description at the UCI Machine Learning repository, and the fact that you can’t take multiple simultaneous readings from the same instrument.

Multi-collinearity



- We used the following chart to assess the correlation of variables with each other. The most highly correlated variables are energy and maxenergy.
- It is also interesting that the gpuls:genergy and gdpuls:gdenergy are somewhat correlated
- For this model, we decided to leave all the main effects variables intact and address any multicollinearity issues after glmnet's automatic feature selection, if necessary (spoiler: it wasn't).

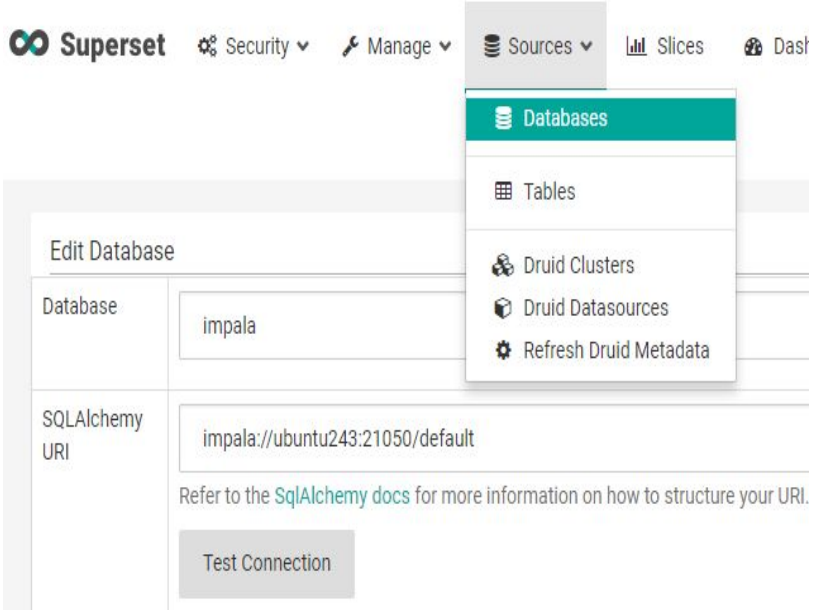
Full Model

- We first built a logistic regression model taking all the observations and variables into account
- Next we predict using this full model and calculate the probability of getting a hazardous bump.
- To convert these probabilities to classes, we defined a threshold.
- A good value for threshold is the mean of the original response variable.
- Probabilities greater than this threshold are categorized into hazardous bump class and probabilities lesser than the threshold are categorized into non-hazardous bump class.
- Model accuracy is then calculated by comparing with the actual class variable. Our calculation show that the logistic regression model with all the observations and variables made a correct prediction 59% of the time.

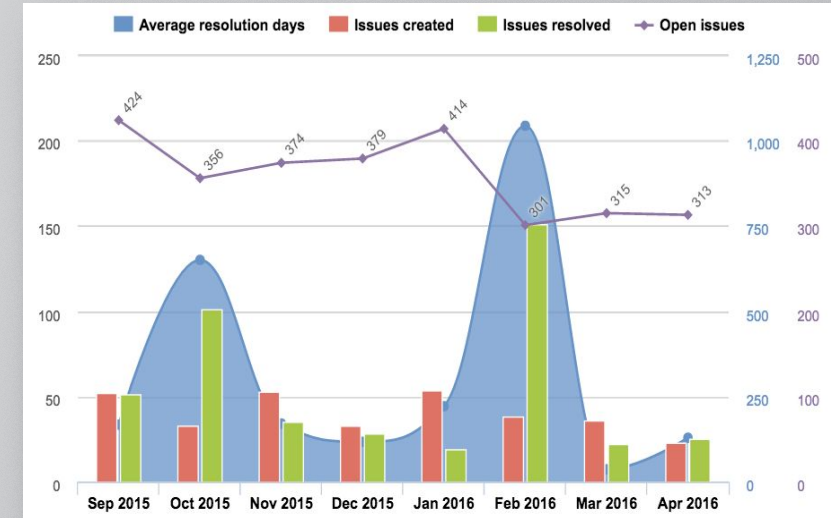
Accuracy using other ML Algorithms

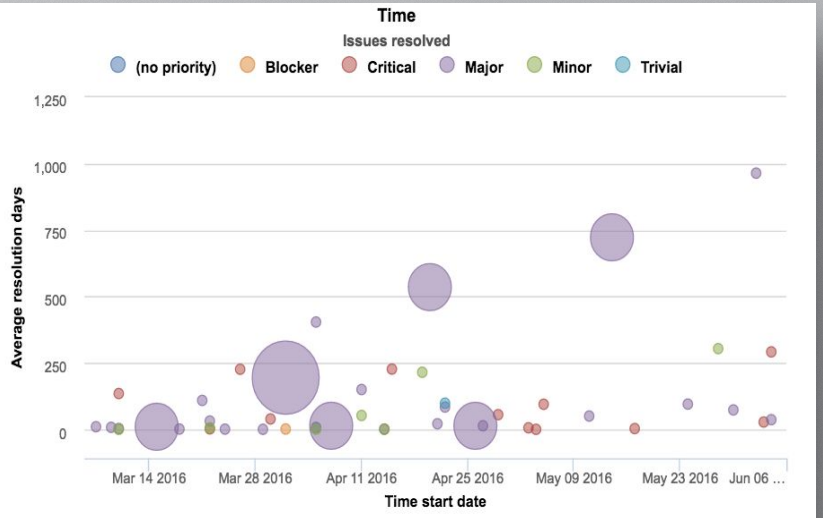
- Decision Tree : 0.8897485493230
- KNN: 0.9323017408123792
- Naive Bayes: 0.11798839458413926,
- Random Forest': 0.922630560928433 3
- LinearSVC': 0.9264990328820116
- Feed Forward NN : 0.912959381044487 4

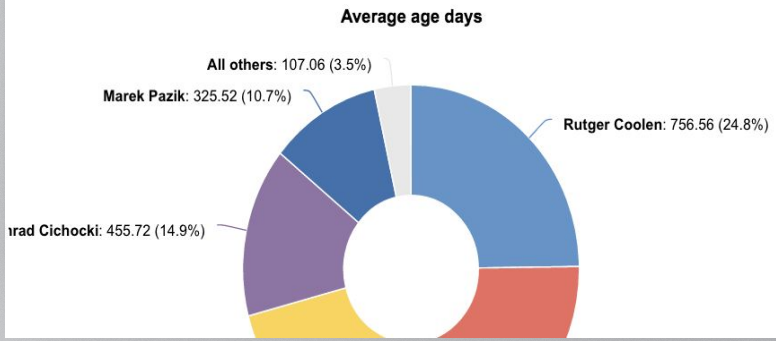
Application In Production : Part - 1 (Data Visualization)

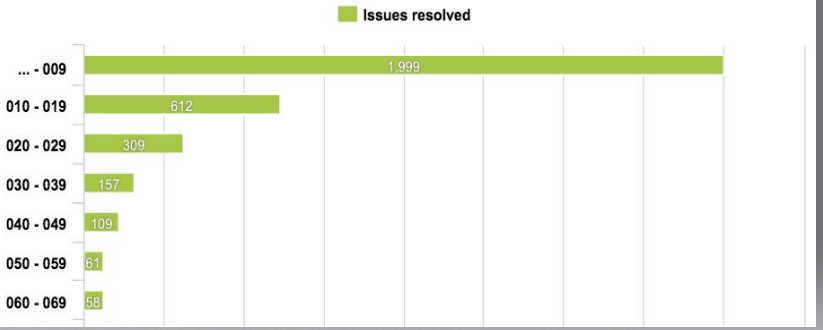


- Writing code for data visualization takes lot of time & effort
- **Solution :**
  - Using Apache Superset
  - Upload Your CSV or Connect to Your Database and Visualize like a pro.
  - Multiple chats are available.

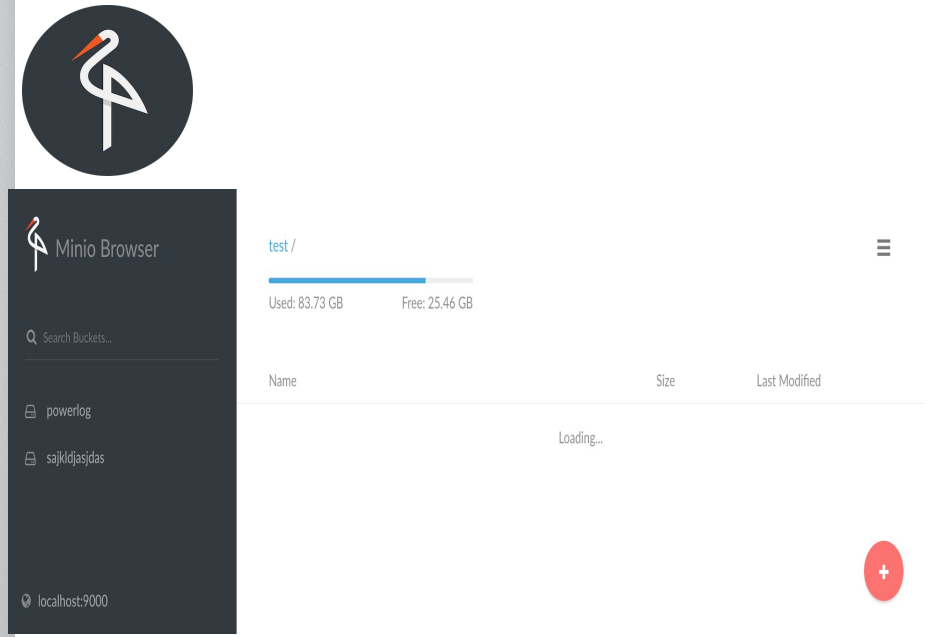

Average resolution days
Issues created
Issues resolved
Open issues
250
200
150
100
50
0
1,250
1,000
750
500
250
0
500
400
300
200
100
0
424
356
374
379
414
301
315
313
Sep 2015
Oct 2015
Nov 2015
Dec 2015
Jan 2016
Feb 2016
Mar 2016
Apr 2016


Time
Issues resolved
(no priority)
Blocker
Critical
Major
Minor
Trivial
Average resolution days
1,250
1,000
750
500
250
0
Mar 14 2016
Mar 28 2016
Apr 11 2016
Apr 25 2016
May 09 2016
May 23 2016
Jun 06 ...
Time start date


Average age days
All others: 107.06 (3.5%)
Marek Pazik: 325.52 (10.7%)
Rutger Coolen: 756.56 (24.8%)
ırad Cichocki: 455.72 (14.9%)


Issues resolved
... - 009
010 - 019
020 - 029
030 - 039
040 - 049
050 - 059
060 - 069
1,999
612
309
157
109
61
58

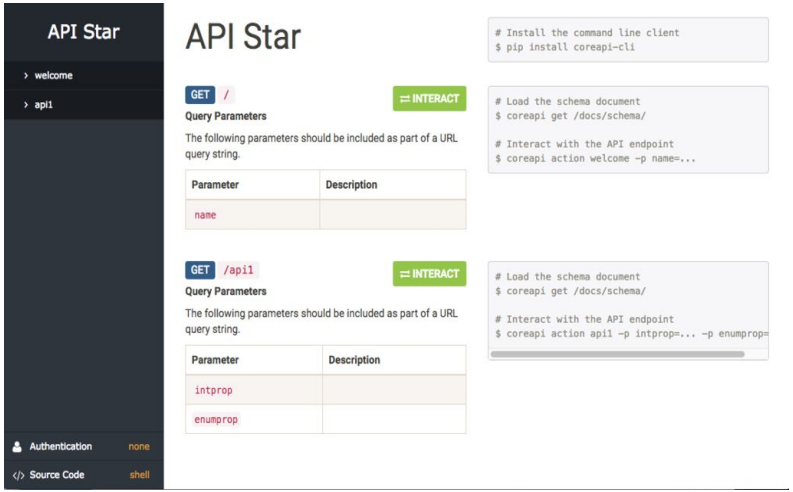
Application In Production : Part - 2 (Building Model)



Minio

- Minio Browser provides minimal set of UI to manage buckets and objects on minio server.
- Having an S3-compatible API means once configured, Minio acts as a gateway to B2.
-

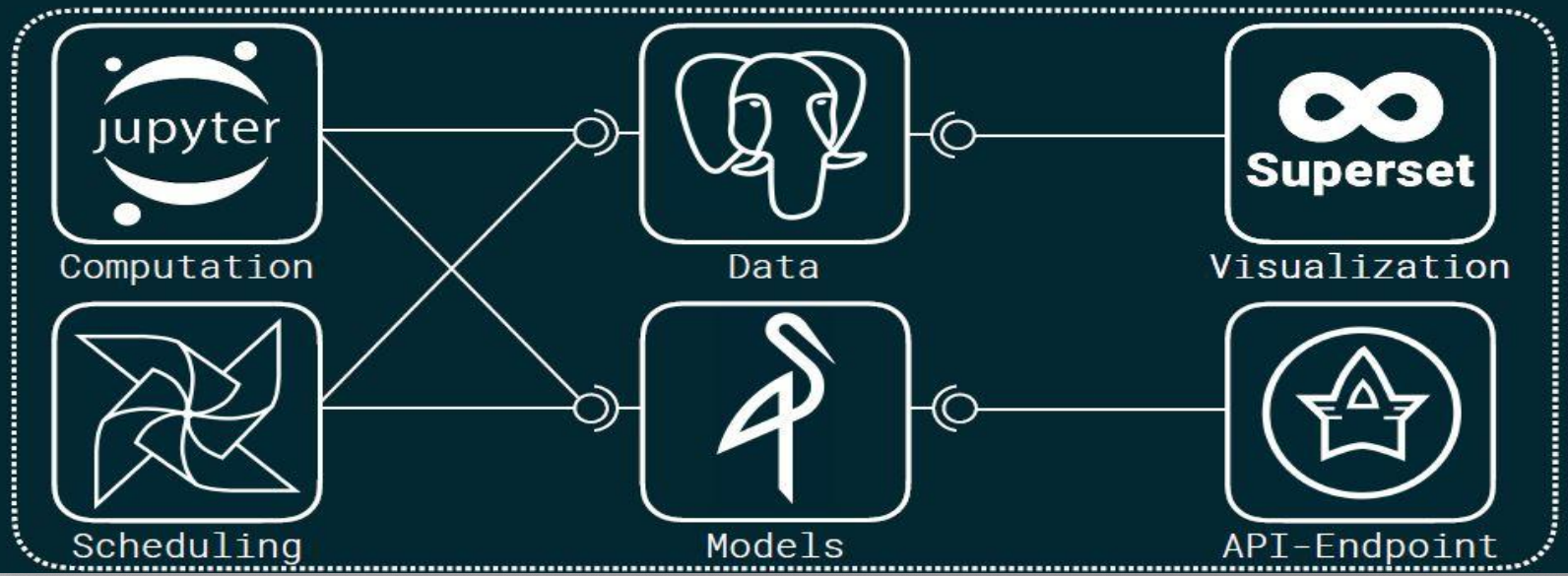
Application In Production : Part - 3 (Predict as a Service)



API Star

- API Star is a toolkit for working with OpenAPI or Swagger schemas.
- No need to write REST-API From scratch.
- Validate API schema documents, and provide contextual errors.
- Validate requests and responses
- Can integrate with frameworks like Django & Flask.

Architecture



Conclusion

- This seismic bump problem is an interesting problem, an important problem, and has a fascinating data set.
- We chose to Randomly Under-Sample, but we could also do Synthetic Minority Over-Sampling, Cluster-Based Over-Sampling, Random Over-Sampling, Algorithmic Ensemble Sampling, Bagging, Boosting, and probably many other techniques which aren't covered in a graduate level statistics course.
- We chose a pretty straightforward solution, which definitely impacts the performance of our model. Our logistic regression model predicts the test set with around 60 to 70% accuracy, after random under-sampling.
- While the methods for calculating the results of these various methods aren't clearly documented in the data set, we can assume that we understand a few of them through inference. Accuracy (Acc.) is the percentage of times our model correctly predicted the class in the test set.

Some different ideas on this dataset:

- Where are the geophones located? Is it possible to do TRM (time reverse modeling) of stress wave propagation? With this dataset we simply know that if a rock burst occurs in a longwall but not where.
- What to do with the results? We may end up knowing if a rock burst is likely to occur or not. Can we use the recorded data to estimate locations and therefore control stress release before the burst?
- Can we extract another set of useful information from the raw data from the geophones (what we got here is heavily processed)?

Future Improvements

- We can use Deep Learning Algorithms like RNN, LSTM models to improve the accuracy.

- Can make a Docker Image of our architecture and serve it as a web app.

Thank You !

Any questions?

Drop a mail at : prabhat.kumar.sahu@suiit.ac.in

You can get the code at my Github repo-
https://github.com/TheCaffeineDev/Seismic-Data-Analysis
Or
https://bit.ly/2Qm16Zi