Weldright

Team ID: 227088
Team Member:
Sam Selvaraj
Pushkar Sawant

Data Analysis

Using Python libraries such as Numpy, Matplotlib, Pandas for statistical inference and analysis.

Model Selection

Different machine learning models were built and tested using the balanced data out of which the model with highest accuracy was selected using the sklearn python library.

Demo

A live demo of our working model on AWS EC2 using a Flask server.

Handling Imbalanced Data

Various techniques were implemented out of which the most effective method was selected.

ROI and TCO calculations

The return on investment was formulated and approximated on the basis of assumptions.

Random Majority Undersampling + Minority Duplication + SMOTE oversampling

- Effective output.
- Meaningful duplication using KNN Algorithm.
- Efficient usage of the dataset and effective handling of the minority population.
- Good F1 Score.

```
No Defect                797748
Tungsten Inclusion         4371
Porosity                   1103
Name: Defect, dtype: int64
```

↓

```
No Defect                200000
Porosity                  22060
Tungsten Inclusion        21855
Name: Defect, dtype: int64
```

↓

```
Tungsten Inclusion       200000
Porosity                 200000
No Defect                200000
Name: Defect, dtype: int64
```

Support Vector Machine

	precision	recall	f1-score	support
0	0.14	0.01	0.01	4301
1	0.45	0.70	0.55	4353
2	0.57	0.80	0.67	4370
accuracy			0.50	13024
macro avg	0.39	0.50	0.41	13024
weighted avg	0.39	0.50	0.41	13024

AdaBoost

	precision	recall	f1-score	support
0	0.52	0.44	0.48	16619
1	0.55	0.56	0.55	16720
2	0.76	0.85	0.80	16661
accuracy			0.62	50000
macro avg	0.61	0.62	0.61	50000
weighted avg	0.61	0.62	0.61	50000

Decision Tree

	precision	recall	f1-score	support
0	1.00	0.99	1.00	164377
1	1.00	1.00	1.00	164263
2	1.00	1.00	1.00	164427
accuracy			1.00	493067
macro avg	1.00	1.00	1.00	493067
weighted avg	1.00	1.00	1.00	493067

Random Forest

	precision	recall	f1-score	support
0	1.00	1.00	1.00	164377
1	1.00	1.00	1.00	164263
2	1.00	1.00	1.00	164427
accuracy			1.00	493067
macro avg	1.00	1.00	1.00	493067
weighted avg	1.00	1.00	1.00	493067

GradientBoost

	precision	recall	f1-score	support
0	0.89	0.73	0.80	3662
1	0.85	0.95	0.90	3642
2	0.89	0.94	0.92	3730
accuracy			0.87	11034
macro avg	0.88	0.87	0.87	11034
weighted avg	0.88	0.87	0.87	11034

Inference :

1.Decision Tree and Random Forest are clearly overfitting the data.

2.SVM and Adaboost is not giving a good performance.(underfitting)

Extreme Gradient Boosting Model(XgBoost)

- The overall accuracy of the model is 94-98%.
- With an acceptable and good individual F1 score.
- In broad terms, it’s the efficiency, accuracy, and feasibility of this algorithm.
- It has both linear model solver and tree learning algorithms. So, what makes it fast is its capacity to do parallel computation on a single machine.
- It also has additional features for doing cross-validation and finding important variables.

	precision	recall	f1-score	support
0	0.97	0.84	0.90	40110
1	0.92	1.00	0.96	39858
2	0.92	0.97	0.95	40032
accuracy			0.94	120000
macro avg	0.94	0.94	0.93	120000
weighted avg	0.94	0.94	0.93	120000

	precision	recall	f1-score	support
0	0.99	0.89	0.94	657661
1	0.95	1.00	0.97	657721
2	0.94	0.99	0.97	656886
accuracy			0.96	1972268
macro avg	0.96	0.96	0.96	1972268
weighted avg	0.96	0.96	0.96	1972268

	precision	recall	f1-score	support
0	1.00	0.94	0.97	657661
1	0.98	1.00	0.99	657721
2	0.96	1.00	0.98	656886
accuracy			0.98	1972268
macro avg	0.98	0.98	0.98	1972268
weighted avg	0.98	0.98	0.98	1972268

Real World Applications:

The XgBoost has also been widely adopted by industry users, including Google, Alibaba and Tencent, and various startup companies.

According to a popular article in Forbes, XgBoost can scale with hundreds of workers (with each worker utilising multiple processors) smoothly and solve machine learning problems involving Terabytes of real-world data.

Return of Investment

Hence, formulating profit per prediction by considering the testing accuracy,

the cost incurred by a false prediction and expected profit.

Z_hat = [P - (1 - A) * e]

where,

Z_hat = Profit per prediction

P = Expected profit

A = Accuracy of the model

e = Cost incurred for every false prediction

Formulation Understanding:

1. (1 - A) gives us the loss generated by our model.
2. Multiplying it with the cost incurred for a false prediction will give us the cost incurred because of the loss of the model.
3. Subtracting it from the profit that can be generated from the product will give us the general profit per prediction on using the model.

ROI Example

Estimate

Let,

P = 100 units

A = .96 (from the model)

e = 150 units (cost incurred by faulty prediction)

Therefore,

Z_hat = 94 units

This means that we can save about 94 units by using the model

Total Cost of Ownership

1. Operating System(OS)

We choose the OS required to run our virtual machine on.

2. Location

We can choose the location where we want to set up our server. A location closer to where the company resides is ideal.

3. Daily load

The daily load depends on how long our EC2 instance will be running daily.

4. EC2 specifications

We can choose the EC2 model according to our hardware requirements.

5. Pricing Strategy

AWS provides various pricing models for the EC2 instance. We can choose any one from them.

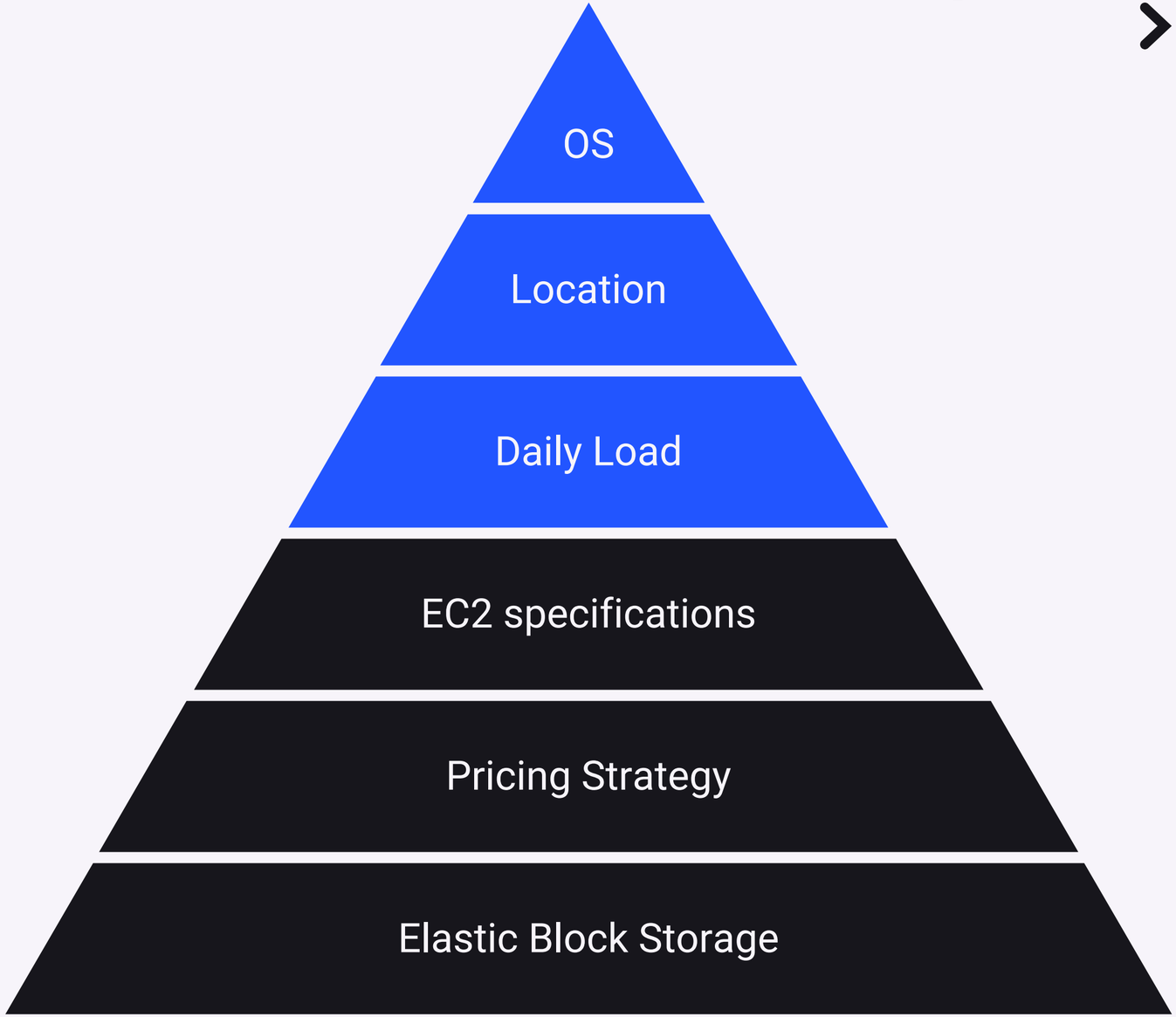
6. Elastic Block Storage

We would require certain storage requirements to be met for our VM. Amazon provides EBS for the storage requirements for our EC2 model.

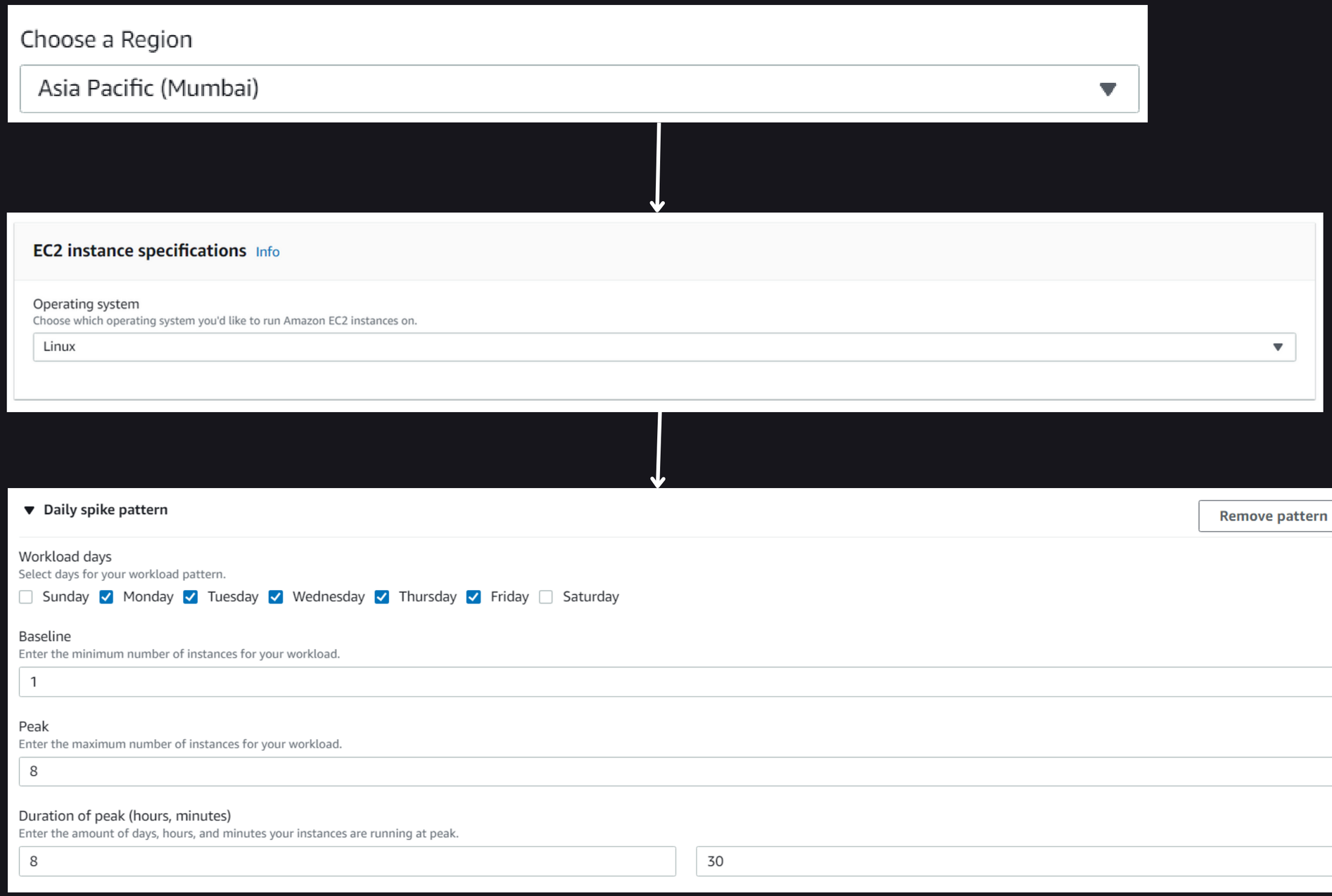
Amazon EC2 On-Demand instances cost (Monthly): 110.14 USD
Amazon EC2 Instance Savings Plans (Monthly): 39.20 USD
Amazon Elastic Block Storage (EBS) total cost (Monthly): 26.05 USD

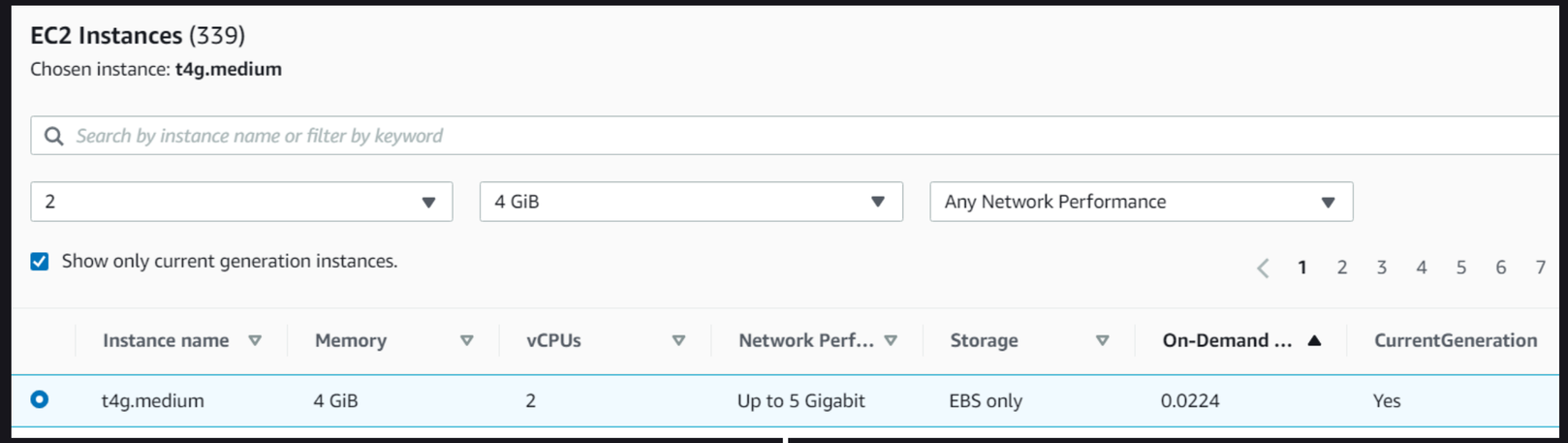
Total Upfront cost: 0.00 USD
Total Monthly cost: 175.39 USD

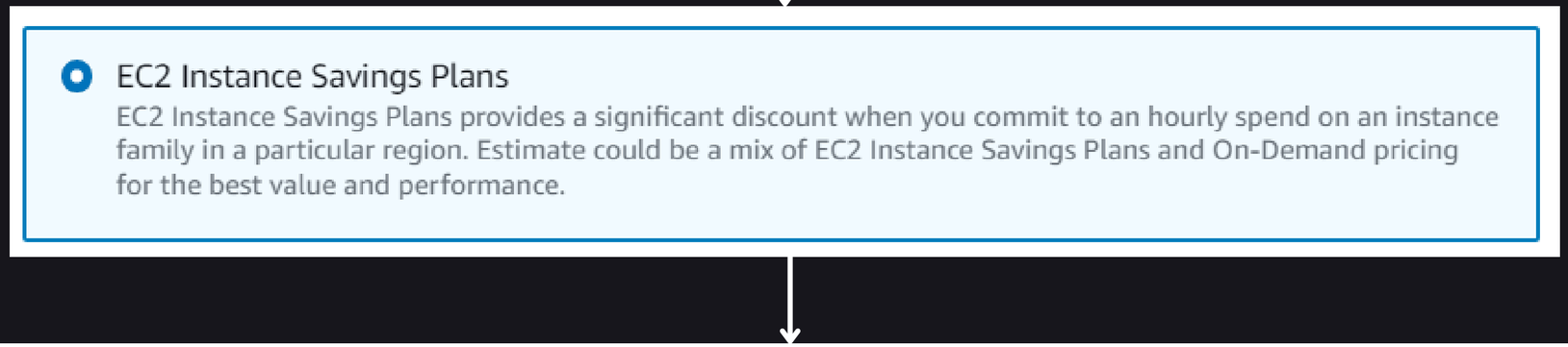
Show Details ▲

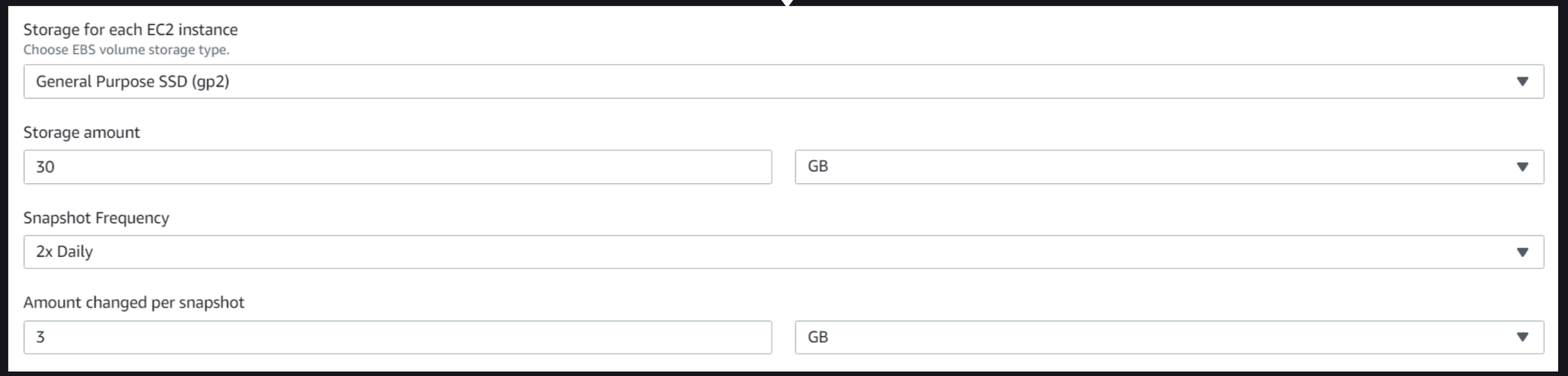


Few snippets of the TCO calculator




EC2 Instances (339)
Chosen instance: t4g.medium
Search by instance name or filter by keyword
2
4 GiB
Any Network Performance
Show only current generation instances.
1 2 3 4 5 6 7
Instance name
Memory
vCPUs
Network Perf...
Storage
On-Demand ...
CurrentGeneration
t4g.medium
4 GiB
2
Up to 5 Gigabit
EBS only
0.0224
Yes

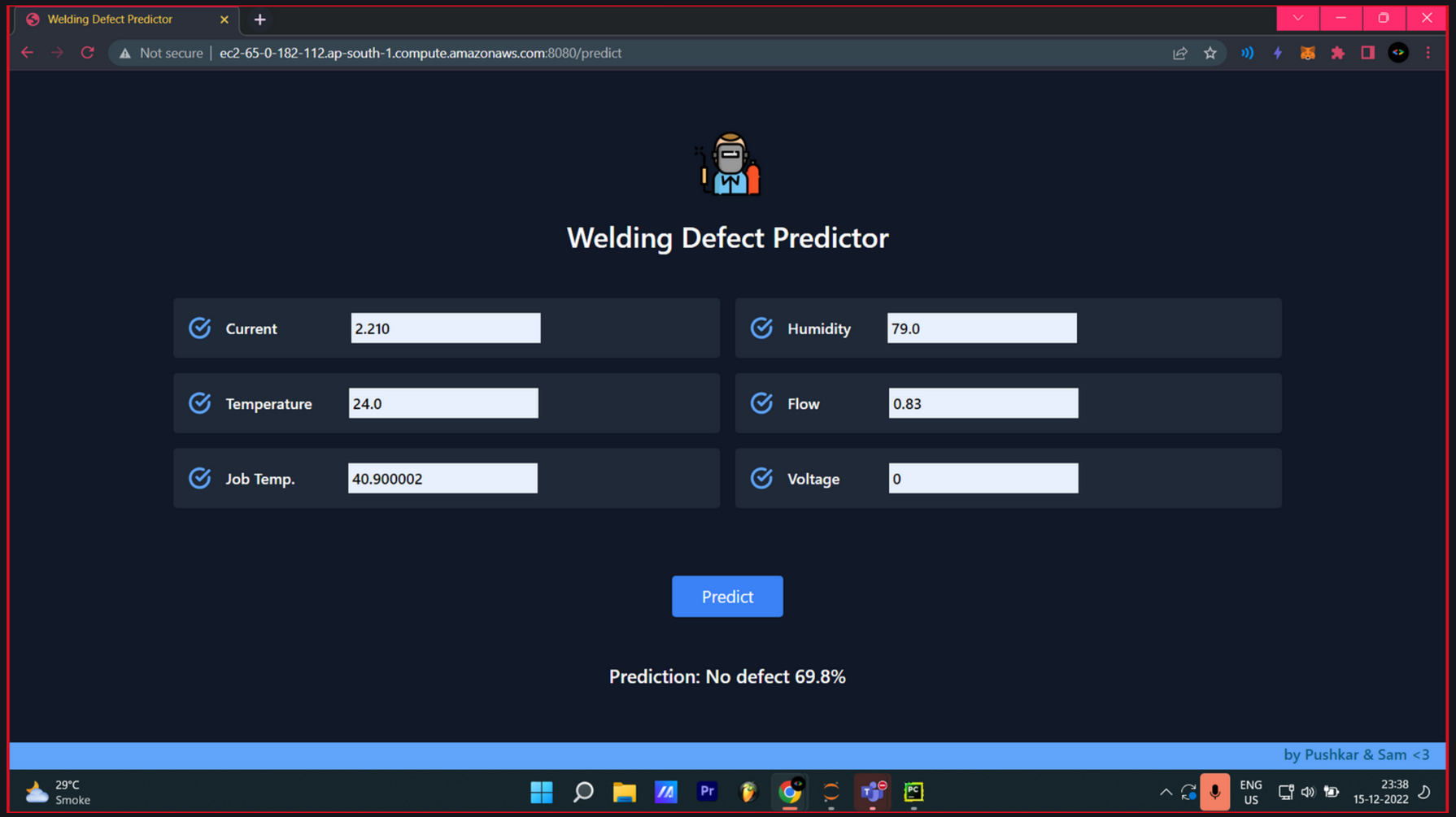

EC2 Instance Savings Plans
EC2 Instance Savings Plans provides a significant discount when you commit to an hourly spend on an instance family in a particular region. Estimate could be a mix of EC2 Instance Savings Plans and On-Demand pricing for the best value and performance.


Storage for each EC2 instance
Choose EBS volume storage type.
General Purpose SSD (gp2)
Storage amount
30
GB
Snapshot Frequency
2x Daily
Amount changed per snapshot
3
GB

Demo

A snippet of our working model on AWS EC2 using a Flask server.

Website Link: https://bit.ly/welding-defect-predictor



Thank You!