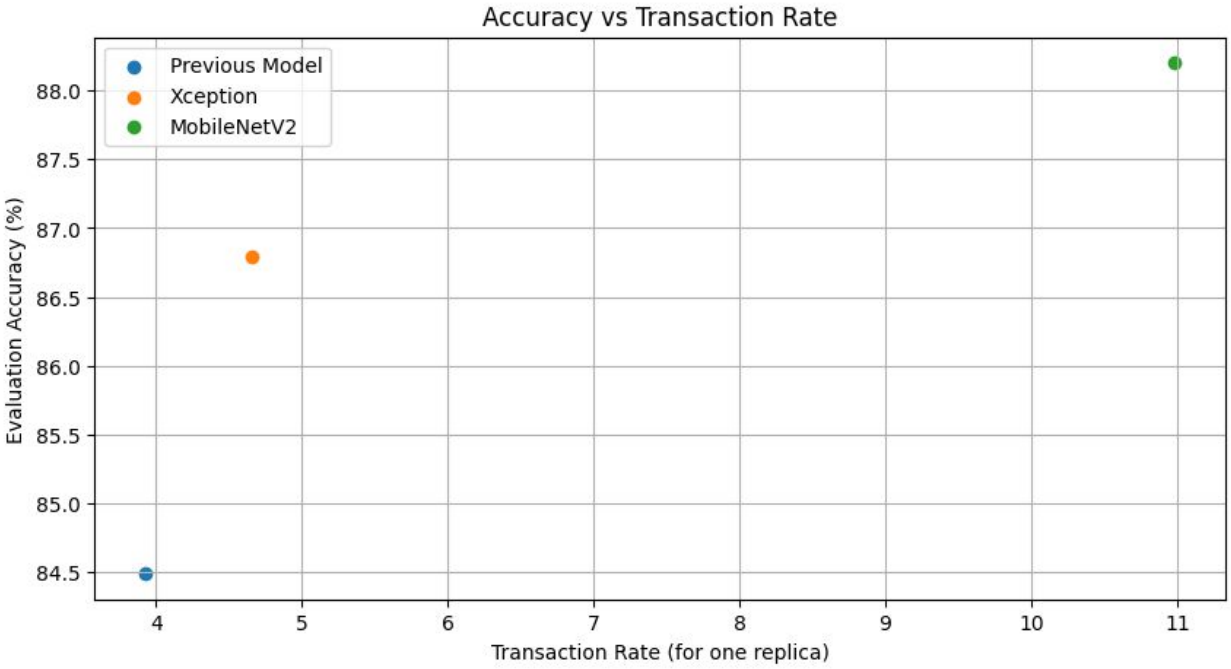
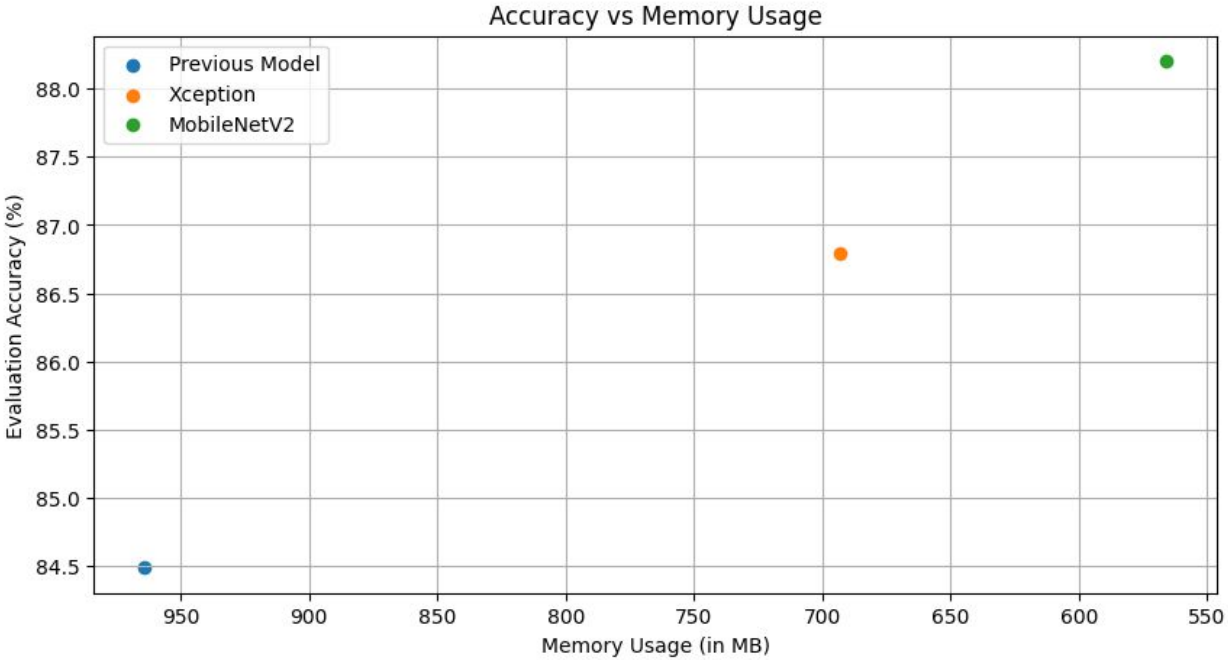
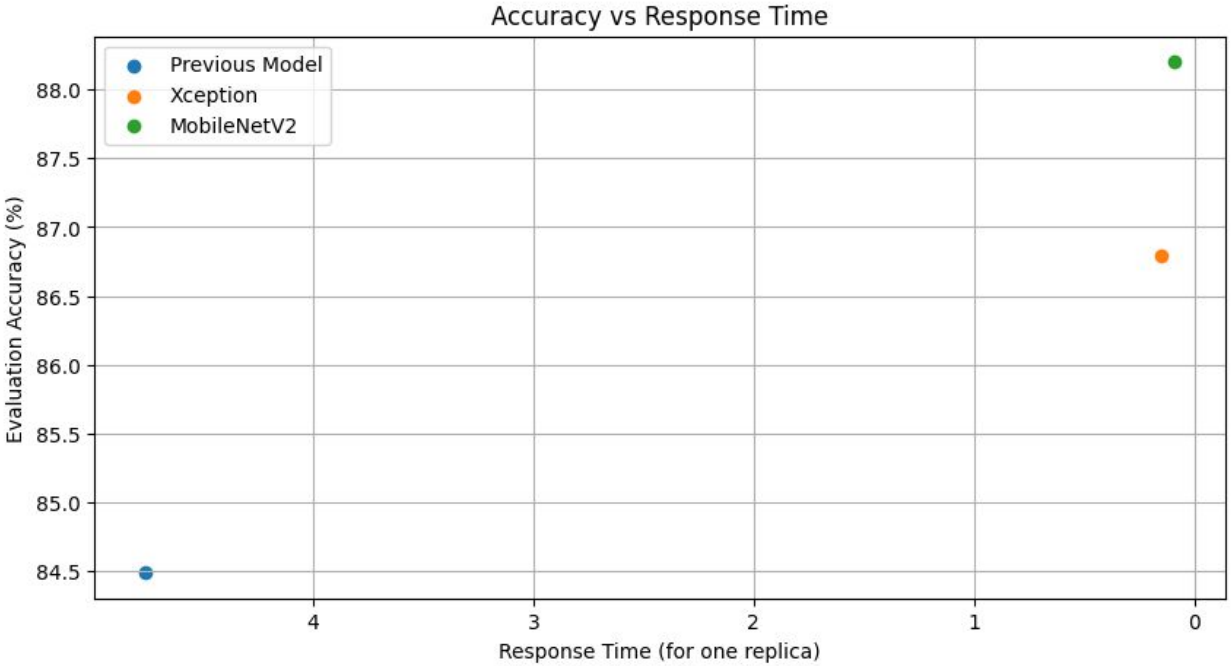
ML Project

Atman Wagle(asw9267)

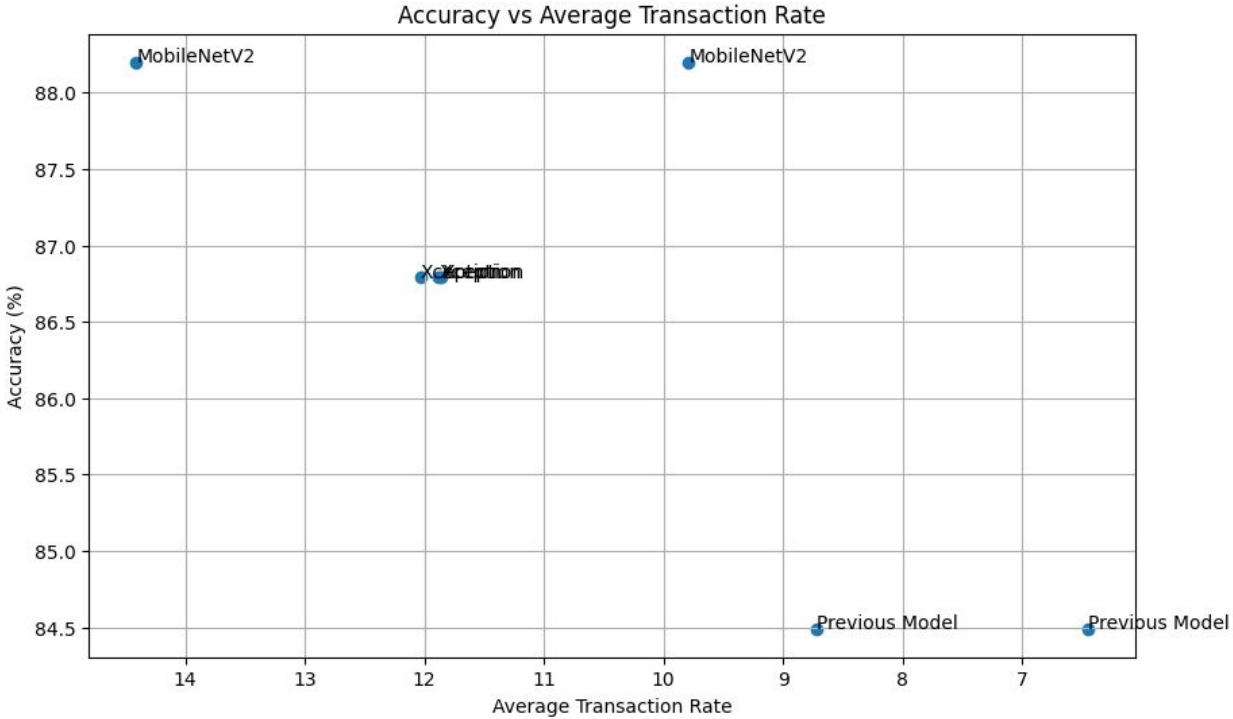
Base Model	Parameters	Training Epochs	Fine Tuning Epochs	**Final Accuracy**	Response Time	Memory Usage	Transactions Per Second	Model Size(MB)
Previous Model	15257419	24	12	84.49	4.76	964 MB	3.93	122.1
Xception	22983219	24	12	86.79	0.15	693 MB	4.66	134.7
MobileNetV2	3590219	24	12	88.2	0.09	566 MB	10.98	28.8

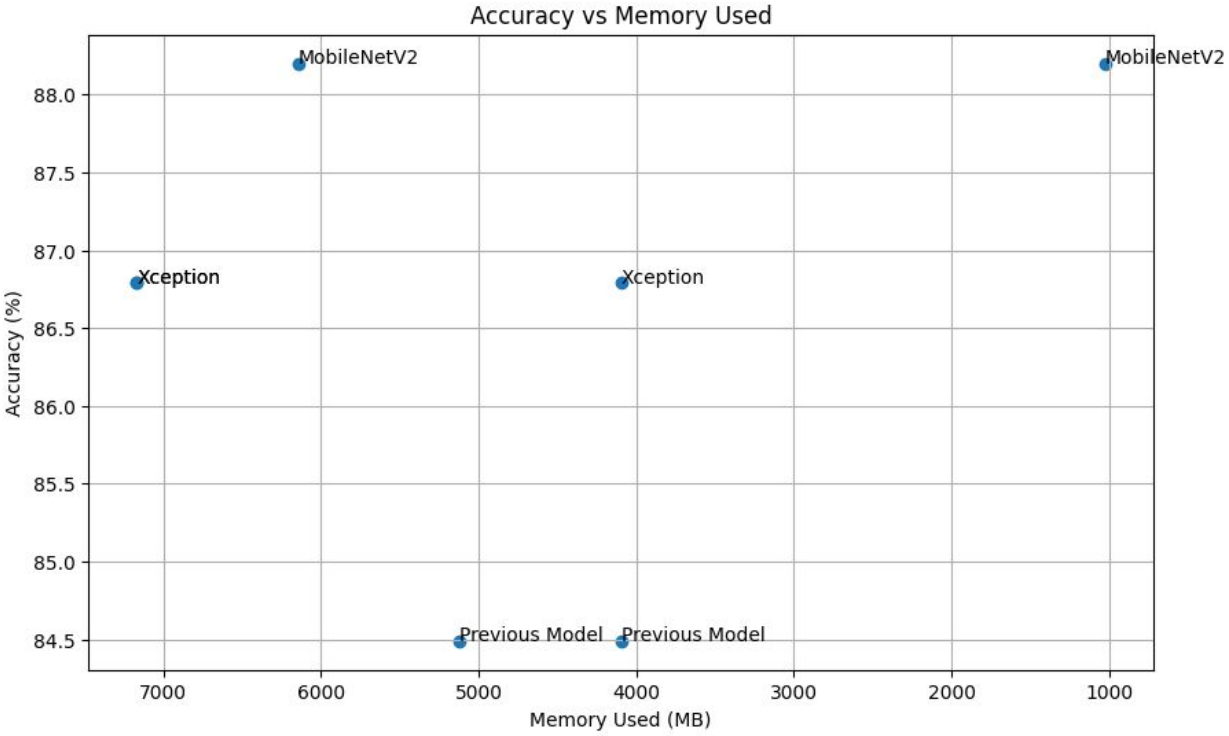

Accuracy vs Transaction Rate
Previous Model
Xception
MobileNetV2
Evaluation Accuracy (%)
88.0
87.5
87.0
86.5
86.0
85.5
85.0
84.5
4
5
6
7
8
9
10
11
Transaction Rate (for one replica)

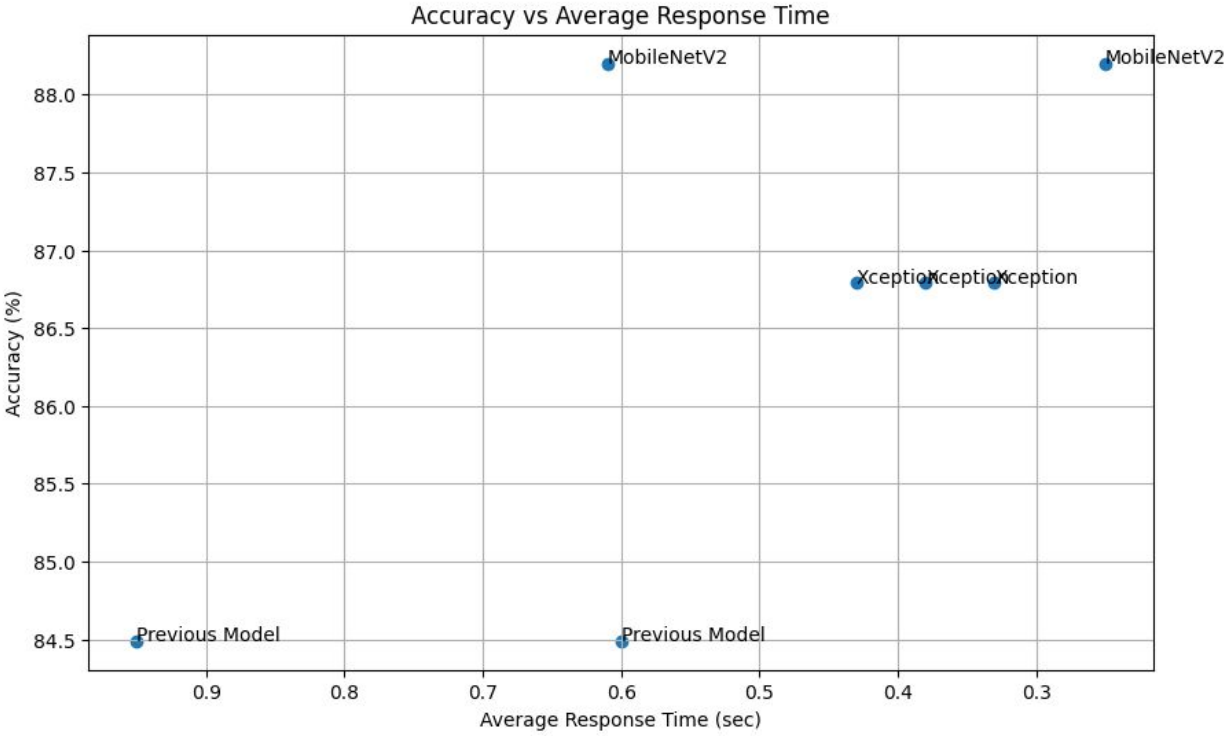

Accuracy vs Memory Usage
Previous Model
Xception
MobileNetV2
Evaluation Accuracy (%)
88.0
87.5
87.0
86.5
86.0
85.5
85.0
84.5
950
900
850
800
750
700
650
600
550
Memory Usage (in MB)


Accuracy vs Response Time
Previous Model
Xception
MobileNetV2
Evaluation Accuracy (%)
88.0
87.5
87.0
86.5
86.0
85.5
85.0
84.5
4
3
2
1
0
Response Time (for one replica)

Model	CPU Limit	RAM Limit(GB)	CPU Utilization Threshold(%)	Average Response Time(sec)	Average Transaction Rate
Previous Model	2	4	40	0.95	6.45
Previous Model	4	5	60	0.6	8.72
Xception	2	4	60	0.43	11.86
Xception	9	7	75	0.33	12.03
Xception	6	7	75	0.38	11.88
MobileNetV2	3	6	75	0.25	14.41
MobileNetV2	1	1	80	0.61	9.79


Accuracy vs Average Transaction Rate
MobileNetV2
MobileNetV2
Previous Model
Previous Model
Accuracy (%)
88.0
87.5
87.0
86.5
86.0
85.5
85.0
84.5
14
13
12
11
10
9
8
7
Average Transaction Rate


Accuracy vs Memory Used
MobileNetV2
MobileNetV2
Xception
Xception
Previous Model
Previous Model
88.0
87.5
87.0
86.5
86.0
85.5
85.0
84.5
Accuracy (%)
7000
6000
5000
4000
3000
2000
1000
Memory Used (MB)


Accuracy vs Average Response Time
MobileNetV2
MobileNetV2
Xception
Xception
Xception
Previous Model
Previous Model
Accuracy (%)
88.0
87.5
87.0
86.5
86.0
85.5
85.0
84.5
0.9
0.8
0.7
0.6
0.5
0.4
0.3
Average Response Time (sec)

- **MobileNetV2** performs 1.5 times faster than the Baseline Model with half the required resources.

- For Systems that are expected to receive a high footfall on their platforms to utilize this ML service, a CPU Limit of 3 and RAM of 6GB is good enough to scale the high number of incoming requests without failure with the **MobileNetV2 model.**

- If there are budget constraints within the organization, 1 CPU and 1 GB RAM would be sufficient to meet this criterion, as it is still better than the original model with more resources, making it a robust cost-effective option to go for.