NYU

Machine Learning (ECE-GY 6143)

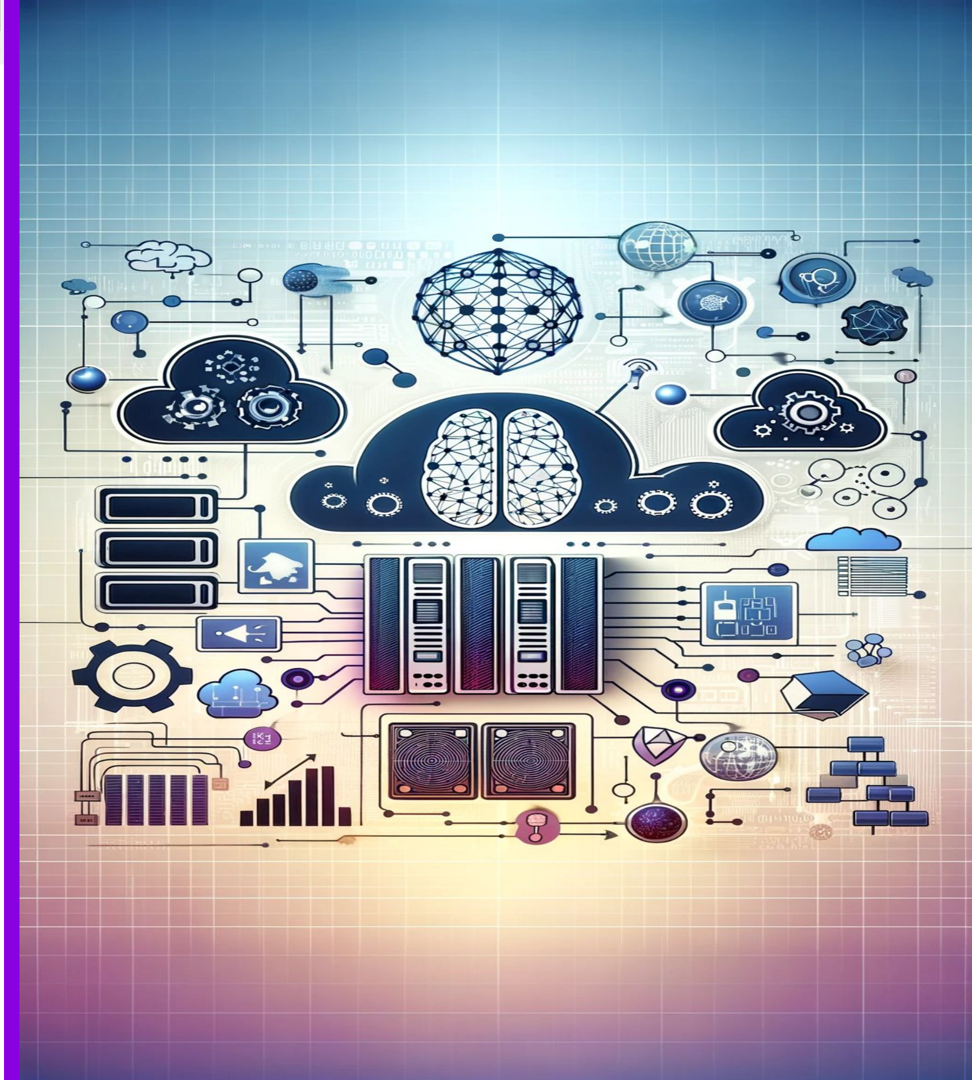
Final Project - Deployment Track

Name: Somya Gupta
Net Id: sg7885@nyu.edu
MS Computer Science
New York University

IMAGINED WITH AI IMAGINED WITH AI

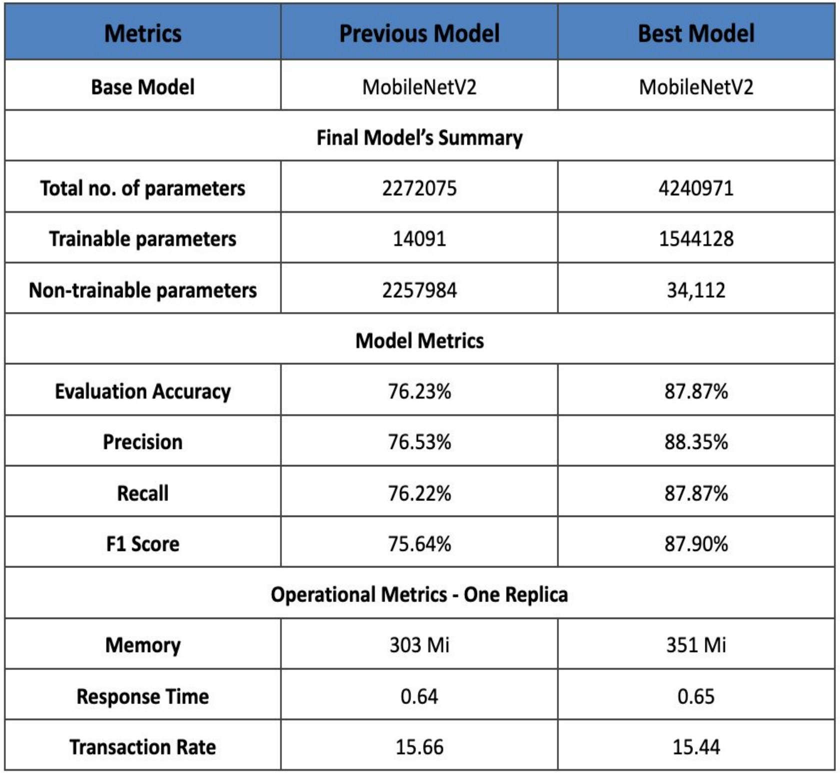
Table of Contents

- Model Summary
- Deployment Options Summary
- Evaluation
- Improvements over Previous Model



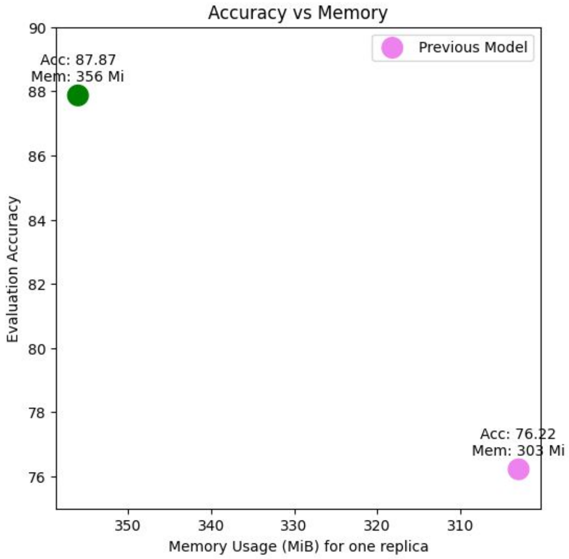
NYU

Model Summary



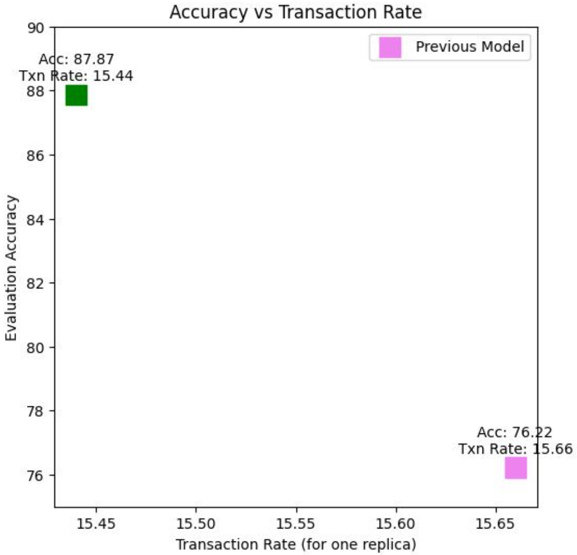
Metrics	Previous Model	Best Model
Base Model	MobileNetV2	MobileNetV2
Final Model's Summary		
Total no. of parameters	2272075	4240971
Trainable parameters	14091	1544128
Non-trainable parameters	2257984	34,112
Model Metrics		
Evaluation Accuracy	76.23%	87.87%
Precision	76.53%	88.35%
Recall	76.22%	87.87%
F1 Score	75.64%	87.90%
Operational Metrics - One Replica		
Memory	303 Mi	351 Mi
Response Time	0.64	0.65
Transaction Rate	15.66	15.44

Accuracy vs Memory - One Replica



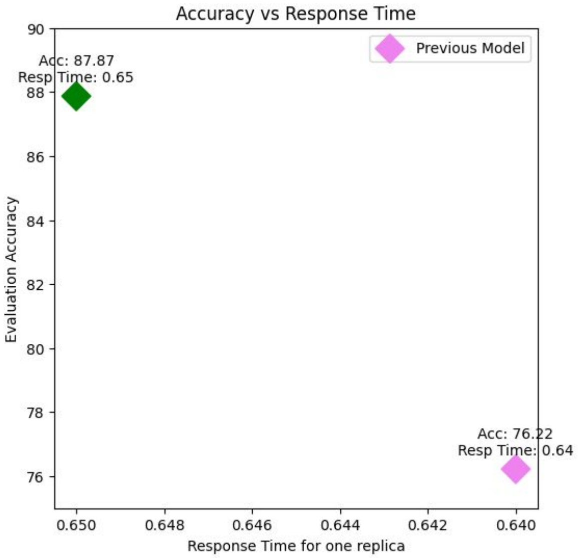
The new model, while more complex, demonstrates operational metrics that are comparable to the older version, with a slight increase in memory usage from 303 Mi to 356 Mi

Accuracy vs Transaction Rate - One Replica



The transaction rate of the best model has a marginal decrease from 15.66 to 15.44 transactions per second.

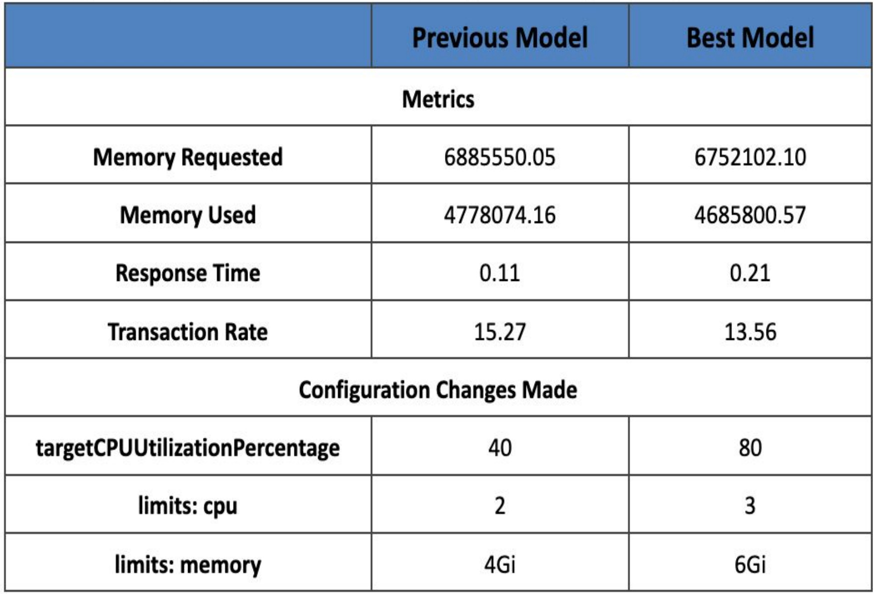
Accuracy vs Response Time - One Replica



The response time of the best model has almost negligibly increased to 0.65 when compared to 0.64 of the previous model.

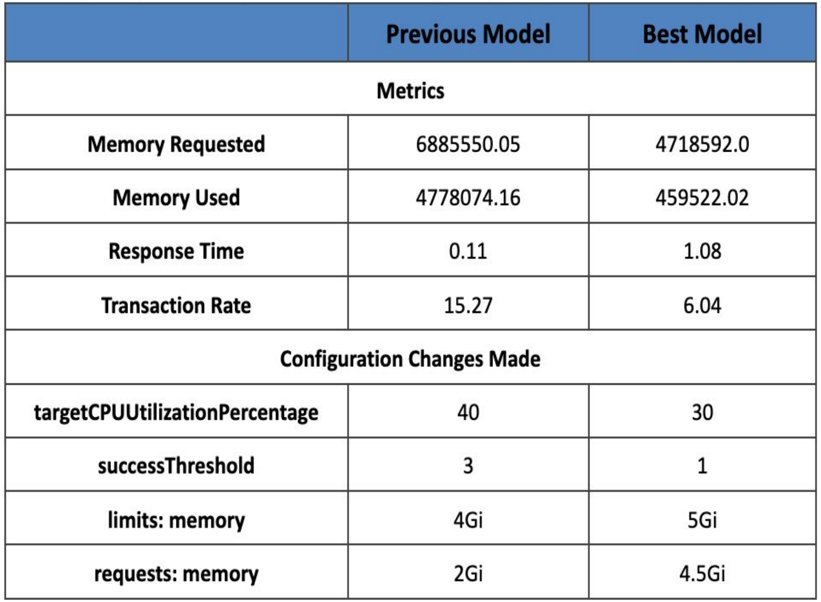
Deployment Options Summary

Strategy 1: Slightly Reduced Memory Usage with Slightly Higher Response Time



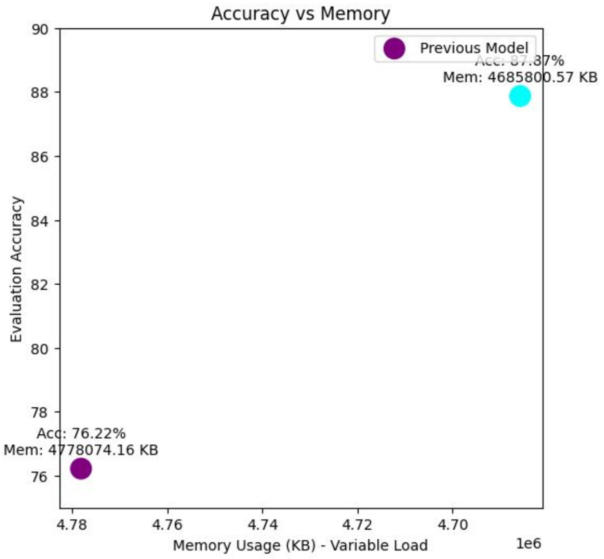
	Previous Model	Best Model
Metrics		
Memory Requested	6885550.05	6752102.10
Memory Used	4778074.16	4685800.57
Response Time	0.11	0.21
Transaction Rate	15.27	13.56
Configuration Changes Made		
targetCPUUtilizationPercentage	40	80
limits: cpu	2	3
limits: memory	4Gi	6Gi

Strategy 2: Significantly Reduced Memory Usage with Greater Response Time



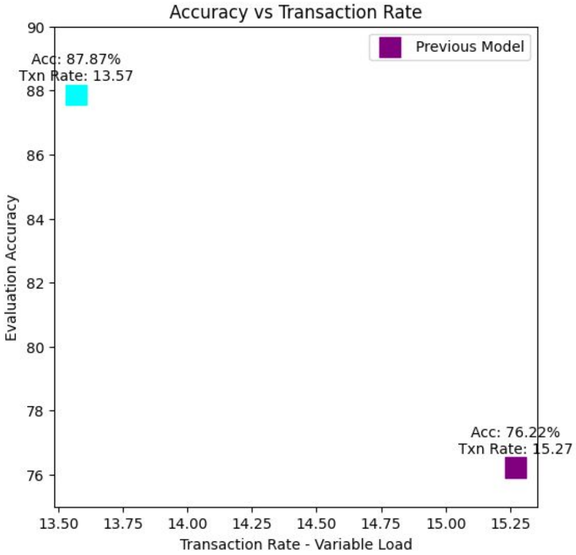
	Previous Model	Best Model
Metrics		
Memory Requested	6885550.05	4718592.0
Memory Used	4778074.16	459522.02
Response Time	0.11	1.08
Transaction Rate	15.27	6.04
Configuration Changes Made		
targetCPUUtilizationPercentage	40	30
successThreshold	3	1
limits: memory	4Gi	5Gi
requests: memory	2Gi	4.5Gi

Accuracy vs Memory - Strategy 1



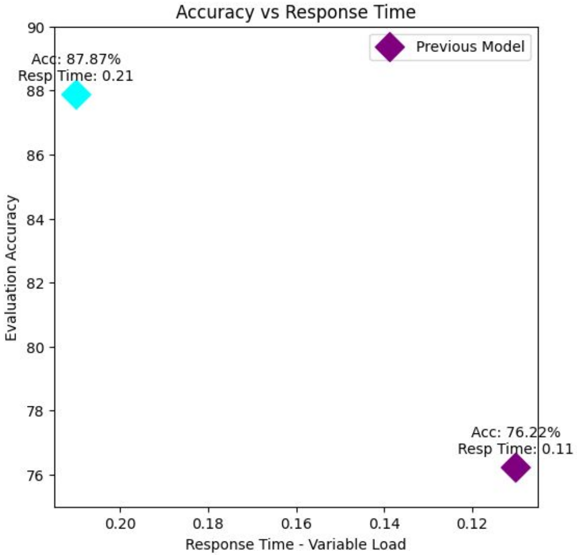
This deployment strategy has a slightly improved memory usage compared to the previous deployment strategy due to increased targetCPUUtilization..

Accuracy vs Transaction Rate - Strategy 1



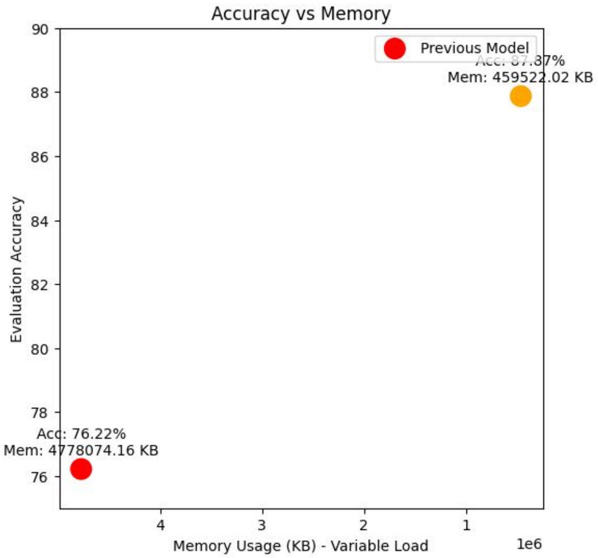
The memory usage efficiency results in a slight drop of transaction rate.

Accuracy vs Response Time - Strategy 1



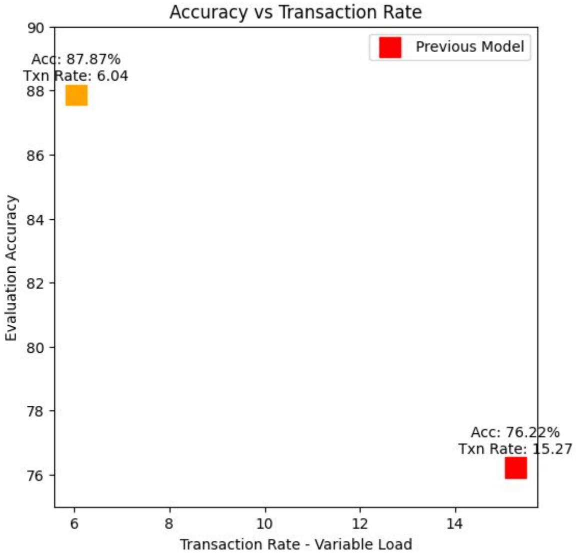
This deployment strategy resulted in a slight increase of response time.

Accuracy vs Memory - Strategy 2



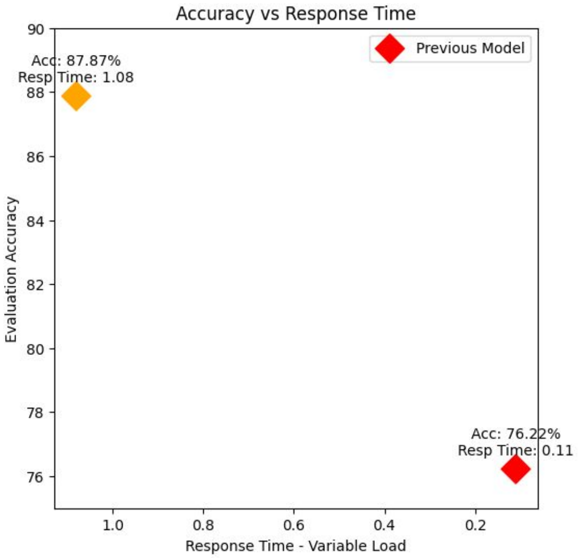
This deployment strategy has a drastic decrease in memory usage when compared to the previous strategy as the memory limit and memory request were increased.

Accuracy vs Transaction Rate - Strategy 2



This deployment strategy saw a great drop in the transaction rate.

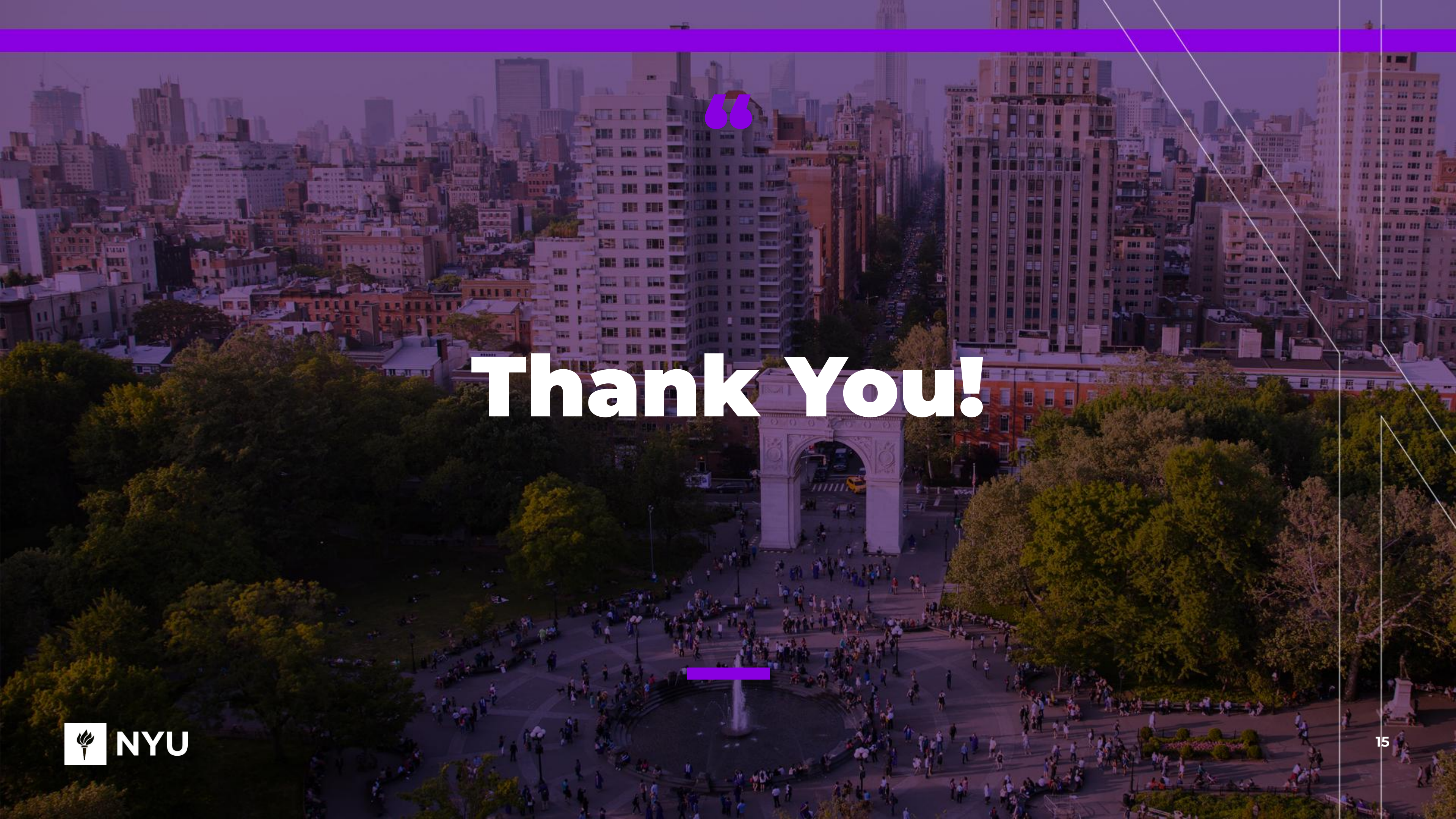
Accuracy vs Response Time - Strategy 2



This deployment strategy witnessed a drastic increase in the response time as memory usage efficiency was given more importance.

Improvements over Previous Model

- **Improved Model Accuracy:** Enhanced from **76.22% to 87.87%**, indicating substantial advancements in classification ability.
- Managers can choose from two strategies:
  - **Strategy 1:** Slightly reduced memory usage, slightly higher response time. This can be used when response time cannot be comprised much.
  - **Strategy 2:** Significantly reduced memory usage, greater response time. This can be used when efficient memory usage is of higher priority than a minimal response time.



Thank You!

NYU