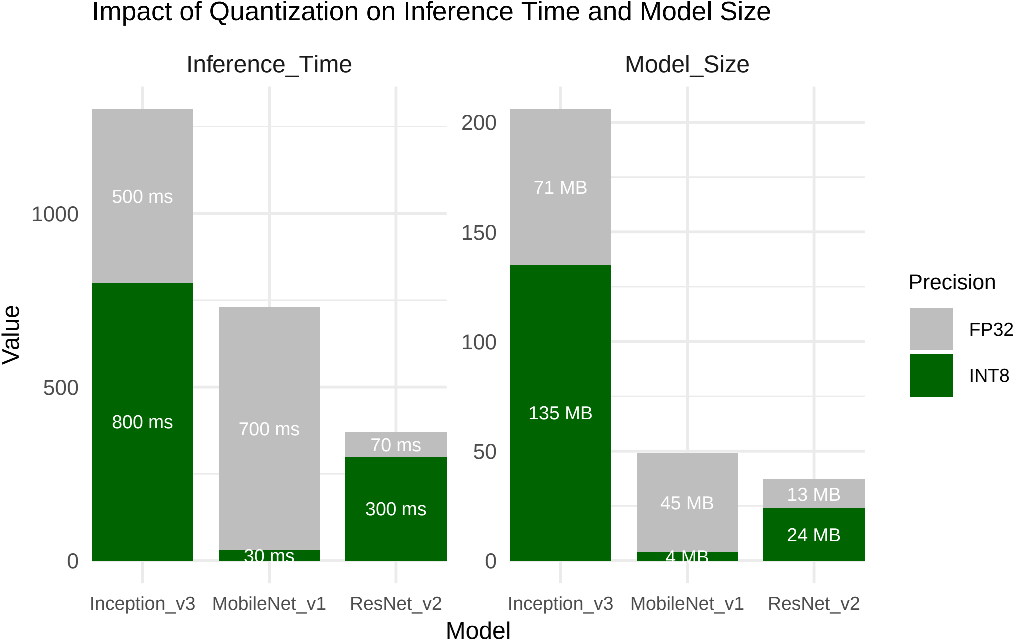
Impact of Quantization on Inference Time and Model Size
Inference_Time
Model_Size
Value
1000
500
0
200
150
100
50
0
500 ms
800 ms
700 ms
30 ms
70 ms
300 ms
71 MB
135 MB
45 MB
4 MB
13 MB
24 MB
Inception_v3
MobileNet_v1
ResNet_v2
Inception_v3
MobileNet_v1
ResNet_v2
Model
Precision
FP32
INT8