Byte Me: A GPU Tale

A SC1015 Mini-Project

Bernard Iskandar, Chen Xinyu, Shao Yingzhan

INTRODUCTION

"MOORE'S LAW IS DEAD"

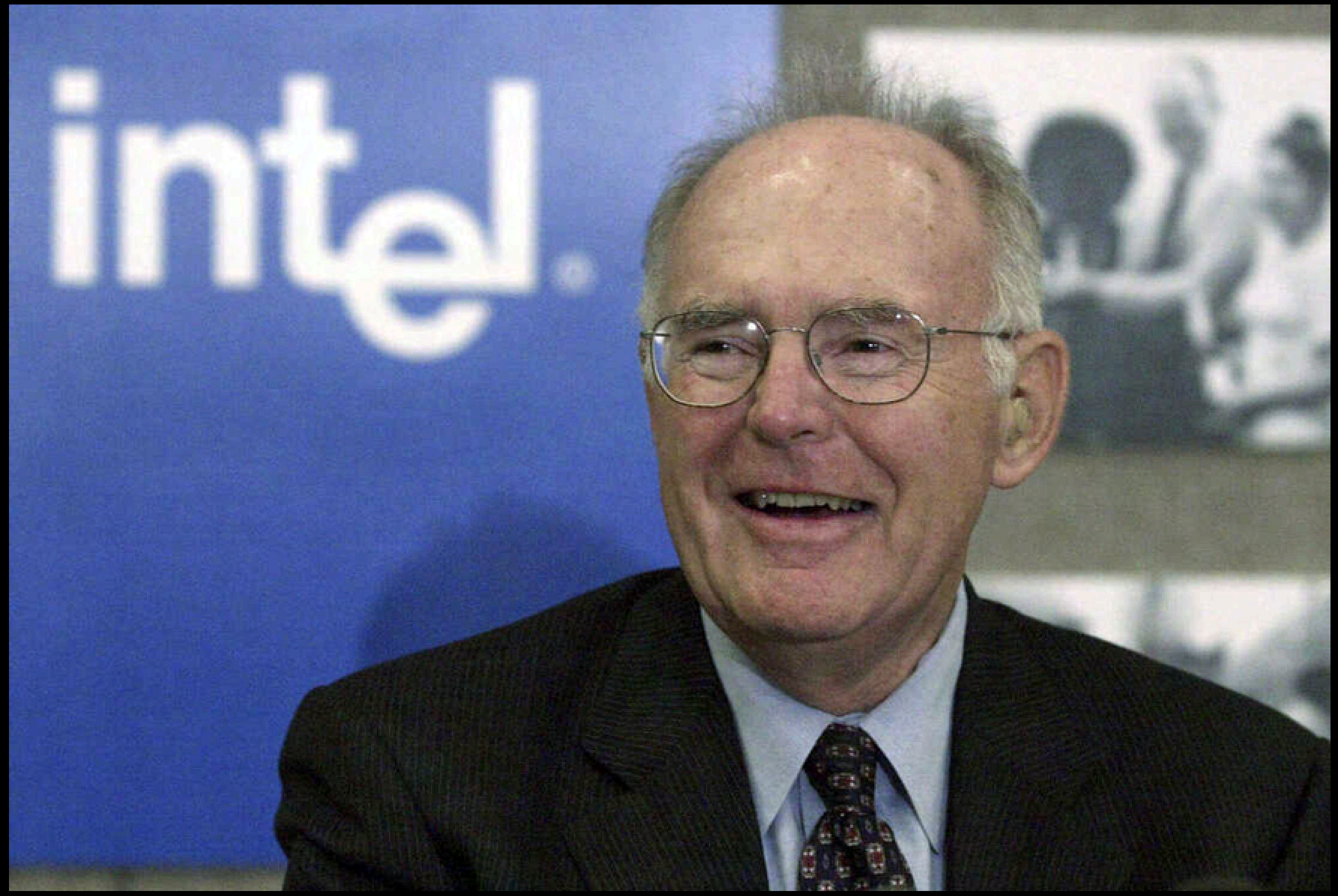
JENSEN HUANG -
CEO,NVIDIA

COMPUTEX 2023

INTRODUCTION

WHAT IS MOORE'S LAW

- Named After Gordon Moore, CEO of Intel
- He observed that the processing power of a chip roughly doubled every 2 years

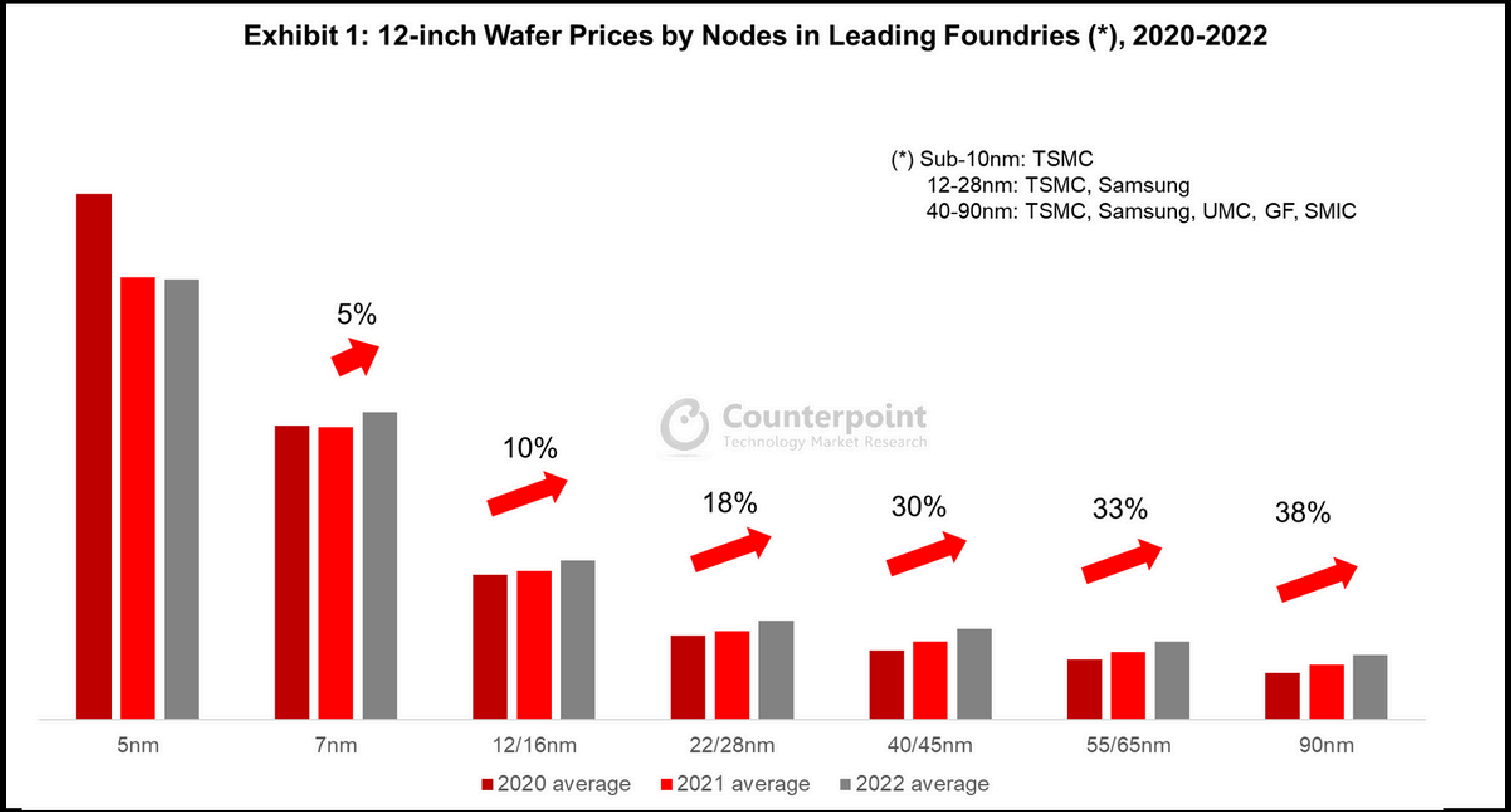


INTRODUCTION

CONTEXT

- As the cost of Semiconductors become more expensive

- There is a need to optimize chip designs



PROBLEM FORMULATION

1. Given a set of chip statistics or features, develop a model to estimate chip performance based on design attributes
2. Identify importance features that affect a chip’s performance

DATA SET

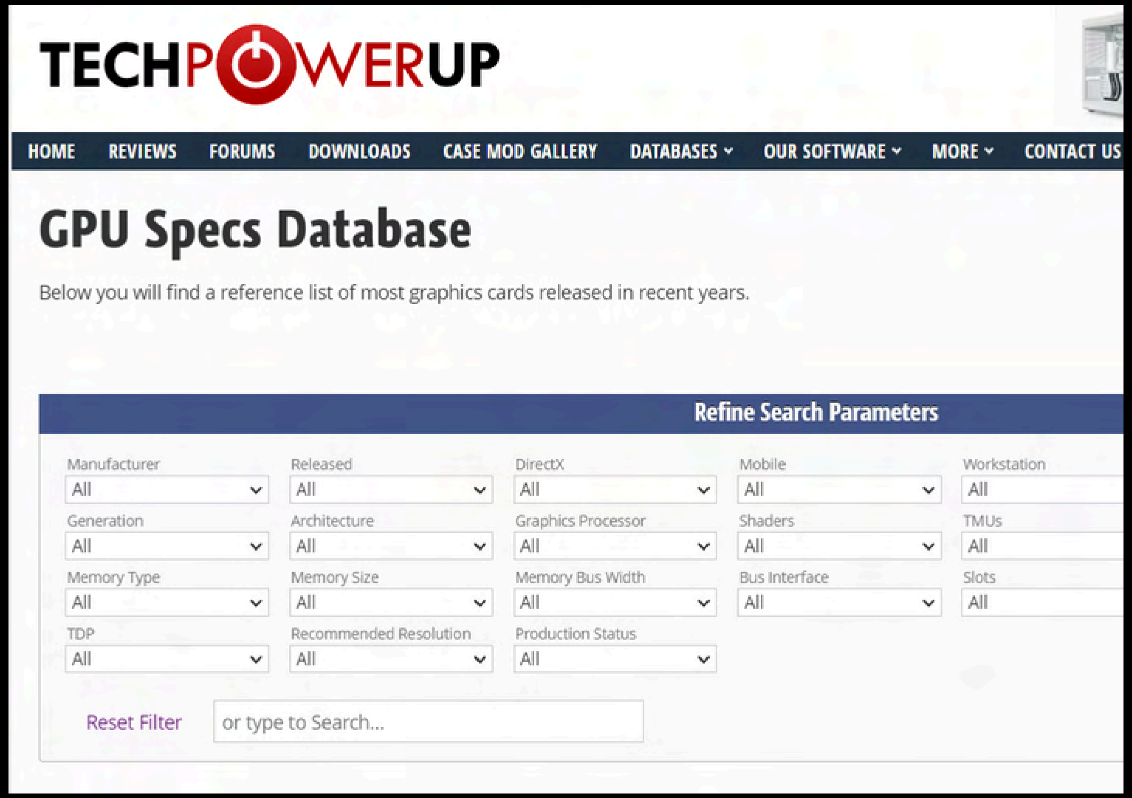
Due to a lack of a ready-made dataset, we had to source our own dataset

Product Name	GPU Chip	Memory	GPU clock	Memory c	Architectu
GeForce G100 OEM	G98S	256 MB, D	540 MHz	400 MHz	Tesla
GeForce GT 120 OEM	G96C	512 MB, D	738 MHz	504 MHz	Tesla
GeForce GT 120 Mac Editior	G96C	512 MB, G	550 MHz	800 MHz	Tesla
GeForce GT 130 OEM	G94B	512 MB, D	500 MHz	500 MHz	Tesla
GeForce GT 130 Mac Editior	G94B	512 MB, G	600 MHz	792 MHz	Tesla
GeForce GT 140 OEM	G94B	1024 MB,	650 MHz	900 MHz	Tesla
GeForce GTS 150 OEM	G92	1024 MB,	738 MHz	1000 MHz	Tesla
GeForce 205 OEM	GT218S	512 MB, D	589 MHz	400 MHz	Tesla 2.0
GeForce 210 PCI	GT216	512 MB, D	475 MHz	400 MHz	Tesla 2.0
GeForce 210 OEM	GT216	1024 MB,	475 MHz	400 MHz	Tesla 2.0
GeForce 210	GT218S	512 MB, D	520 MHz	400 MHz	Tesla 2.0
GeForce 210 Rev. 2	GT218S	1024 MB,	520 MHz	400 MHz	Tesla 2.0
GeForce G210 OEM	G96C	512 MB, D	550 MHz	504 MHz	Tesla
GeForce G210 OEM Rev. 2	GT218S	128 MB, D	589 MHz	400 MHz	Tesla 2.0
GeForce GT 220 OEM	GT215	512 MB, G	506 MHz	700 MHz	Tesla 2.0
GeForce GT 220	GT216	512 MB, D	615 MHz	1000 MHz	Tesla 2.0
GeForce GT 220	G94	1024 MB,	600 MHz	700 MHz	Tesla
GeForce GT 230 OEM	G92B	1536 MB,	500 MHz	500 MHz	Tesla
GeForce GT 230	G94B	512 MB, G	650 MHz	900 MHz	Tesla
GeForce GTS 240 OEM	G92B	1024 MB,	675 MHz	1100 MHz	Tesla
GeForce GT 240	GT215	1024 MB,	550 MHz	850 MHz	Tesla 2.0
GeForce GTS 250	G92B	1024 MB,	675 MHz	1008 MHz	Tesla
GeForce GTS 250	G92B	1024 MB,	702 MHz	1000 MHz	Tesla
GeForce GTX 260 OEM	GT200	1792 MB,	518 MHz	1008 MHz	Tesla 2.0
GeForce GTX 260	GT200	896 MB, G	576 MHz	999 MHz	Tesla 2.0
GeForce GTX 260 Rev. 2	GT200B	896 MB, G	576 MHz	999 MHz	Tesla 2.0
GeForce GTX 260 Core 216	GT200	896 MB, G	576 MHz	999 MHz	Tesla 2.0

DATASET SOURCING

Data Source

- www.techpowerup.com



Webcrawler

- Scraped the Database for data

```python
# Iterate over the DataFrame and scrape details for each GPU
print(f'the start iteration is {iteration}')
print(f'the start index is {start_index}')
for index, row in gpu_data.loc[start_index:].iterrows():
    iteration += 1
    print(f"the iteration is {iteration}")
    time.sleep(random.randint(1,3)) ## to prevent the website from blocking the request
    chip_url = row['Chip URL']
    gpu_url = row['GPU URL']
    chip_details = scrape_gpu__chip_details(chip_url)
    gpu_details = scrape_gpu__perf_details(gpu_url)

    if not chip_details["Architecture"]: # check f the HTTP req failed
        updated_file_path = f'Updated_GPU_Dataset_{iteration}_AMD.xlsx' ## saves the work
        gpu_data.to_excel(updated_file_path, index=False)
        print(f"Updated dataset saved to {updated_file_path}")
        start_index = iteration-1 ##update the start index
        iteration = iteration-1 ## update the iteration to last full iteration
        print(f"the start index is {start_index}")
        raise SystemExit("Stopping execution of this cell.")
```

DATA CLEANING

- Due to the data being in GFLOPS and TFLOP, we converted all to TFLOPS

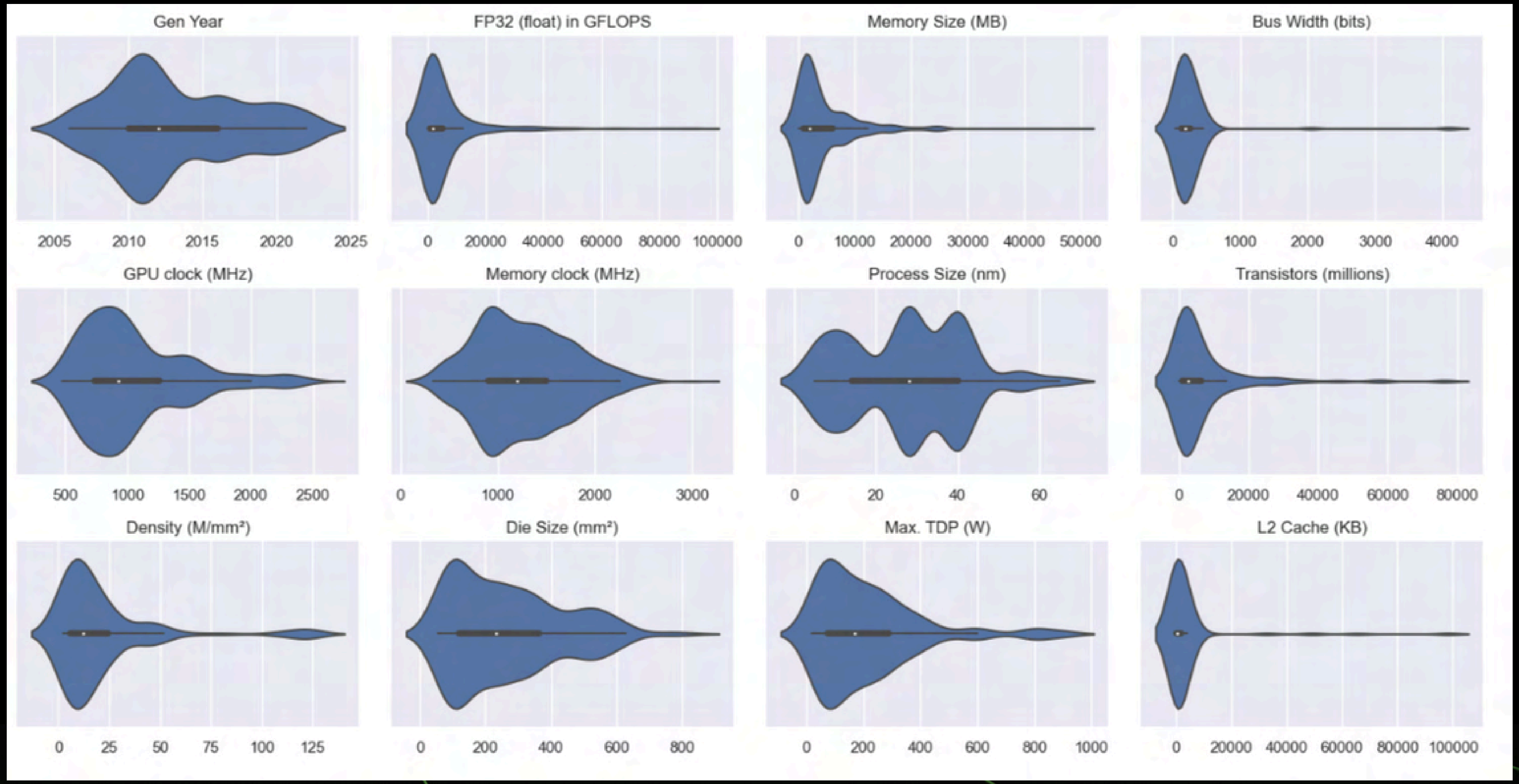
```
def convert_flops(value):
    value = value.replace(",", "")  # Remove commas
    if 'TFLOPS' in value:
        return float(value.replace('TFLOPS', '').strip()) * 1000
    elif 'GFLOPS' in value:
        return float(value.replace('GFLOPS', '').strip())
    else:
        return None  # Just in case there are other formats we haven't considered

# Apply the conversion to the 'FP32 (float)' column
AMD_data['FP32 (float) in GFLOPS'] = AMD_data['FP32 (float)'].apply(convert_flops)

# Cleaning the 'Memory' column by splitting into size, type, and bus width
AMD_data[['Memory Size', 'Memory Type', 'Bus Width']] = AMD_data['Memory'].str.extract(r'(\d+ GB|\c
```

- We also dropped columns that were not needed such as number of Ray tracing cores

8	Bus Width (bits)	218 non-null	int64
9	GPU clock (MHz)	218 non-null	float64
10	Memory clock (MHz)	218 non-null	float64
11	Process Size (nm)	218 non-null	float64
12	Transistors (millions)	218 non-null	float64
13	Density (M/mm²)	218 non-null	float64
14	Die Size (mm²)	218 non-null	float64
15	Max. TDP (W)	218 non-null	float64
16	Pixel Rate (GPixel/s)	218 non-null	float64
17	Texture Rate (GTexel/s)	218 non-null	float64

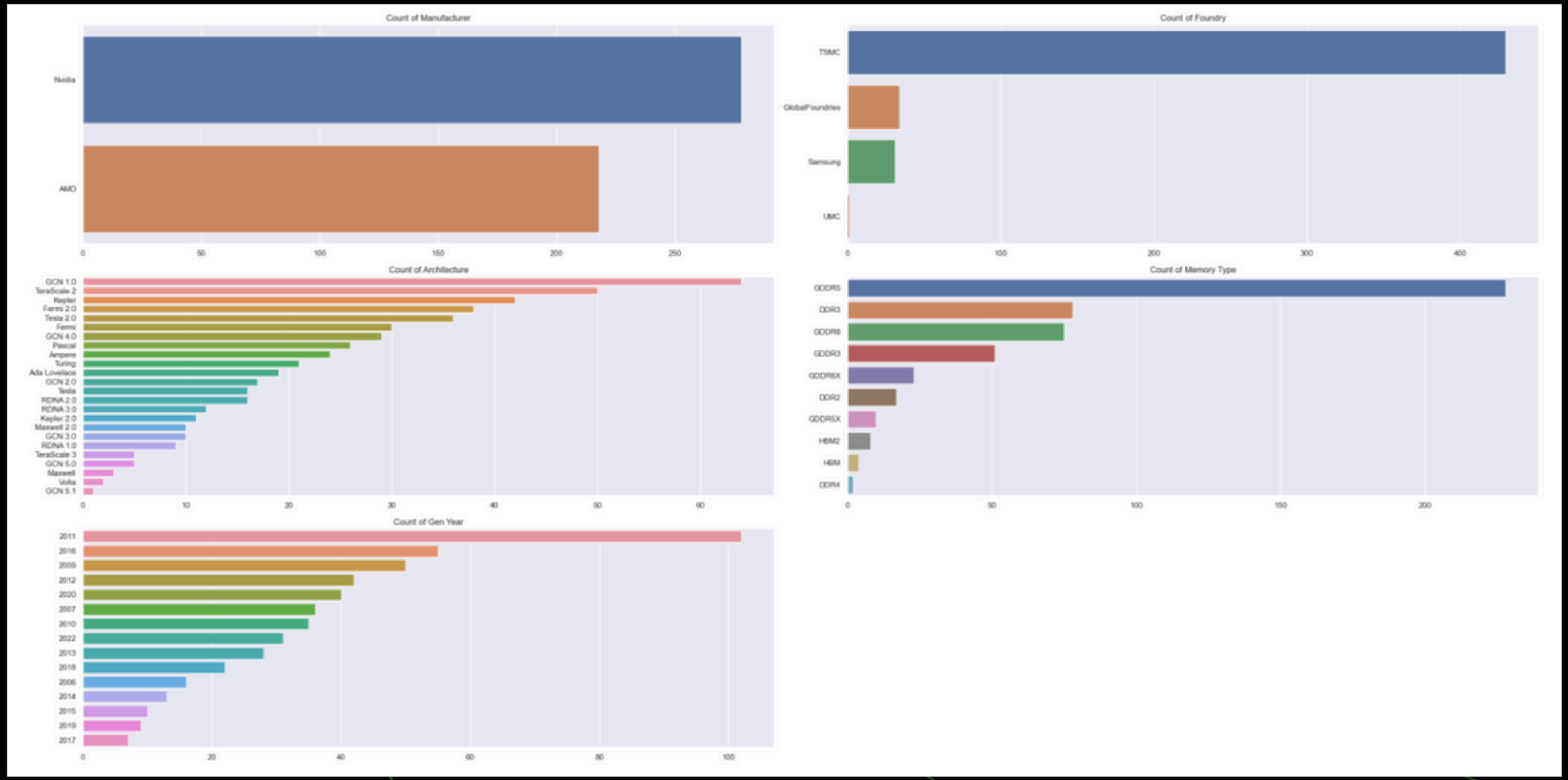
EXPLORATORY DATA ANALYSIS

- Majority of violin plots shows a left skew
- Gen Year later treated as a categorical variable instead
- 11 numerical variables.
- Gen Year later treated as a categorical variable instead



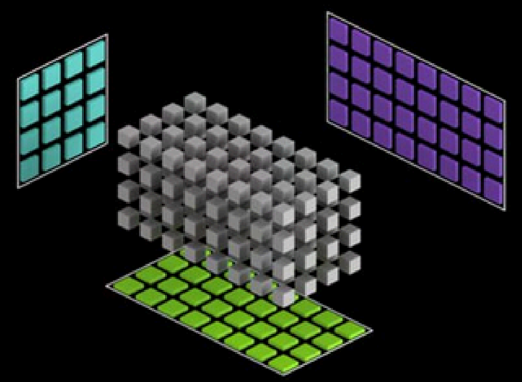
EXPLORATORY DATA ANALYSIS

- 5 categorical variables



THE RESPONSE VARIABLE

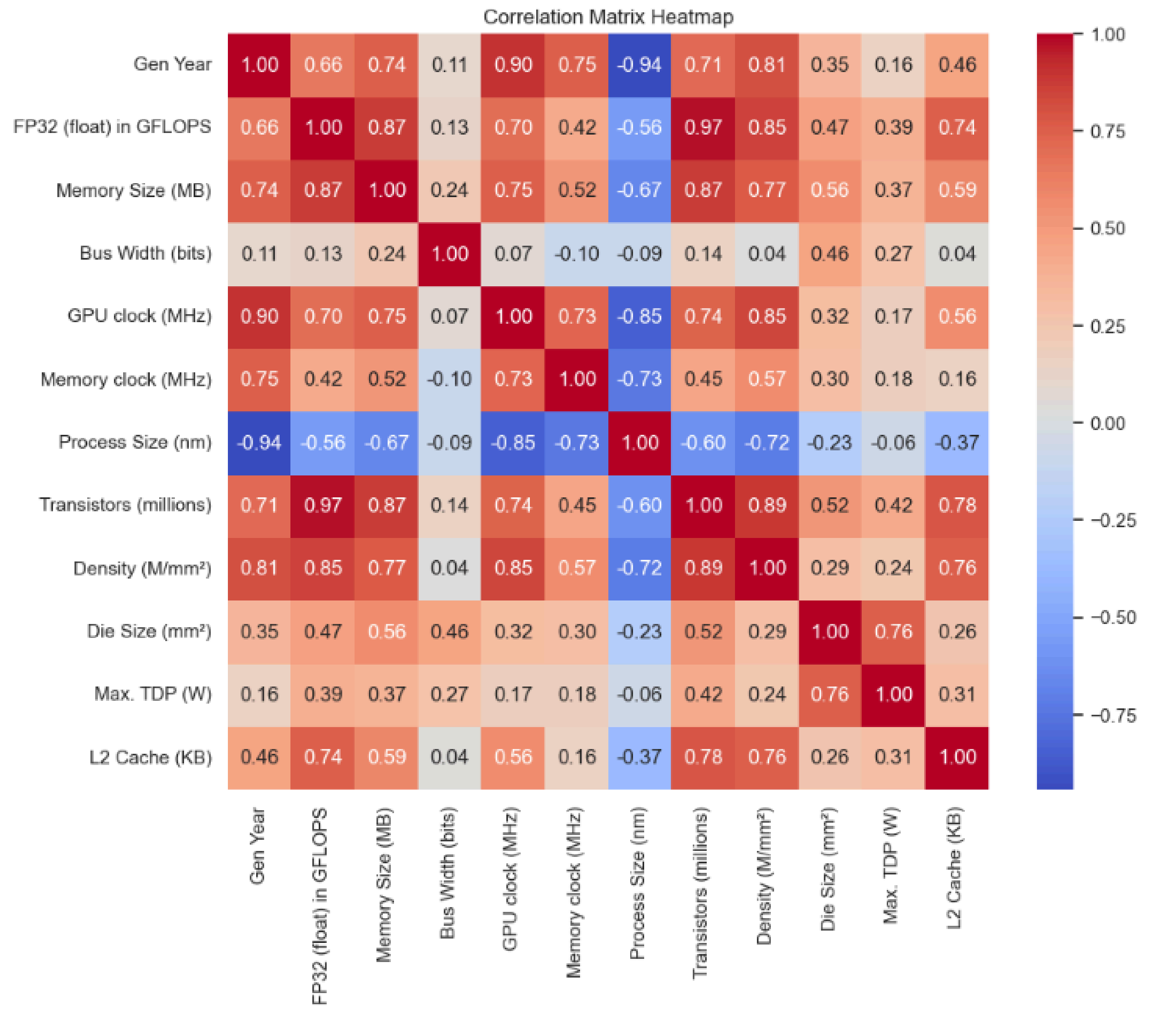
NVIDIA V100 FP32



FP32 in GFlops will be the response variable that measures chip performance.

EXPLORATORY DATA ANALYSIS

- Strong positive correlations between Transistors and FP32 (float) in GFLOPS
- Strong positive correlation between Memory Size (MB) and FP32 (float) in GFLOPS
- Memory Clock (MHz) and Process Size (nm) have weaker correlations with most of the other variables.
- Process Size (nm) has a moderately negative correlation with many variables, such as FP32 (float) in GFLOPS, Transistors, and Texture Rate.



OUR MACHINE LEARNING MODELS

Random Forest

XGBoost

Gradient-Boosting

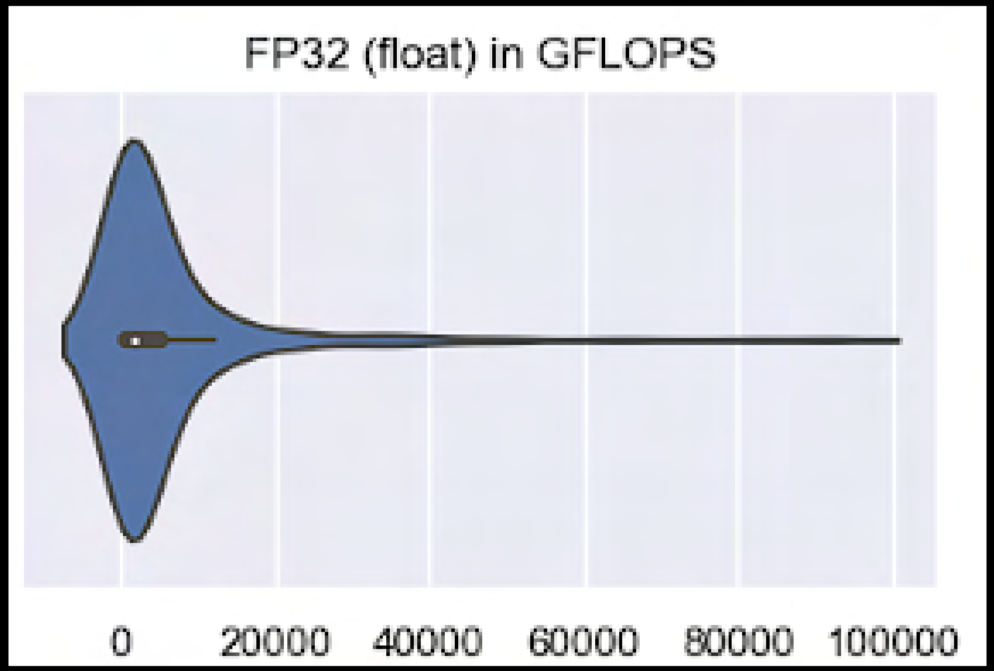
GOODNESS OF FIT COMPARISON

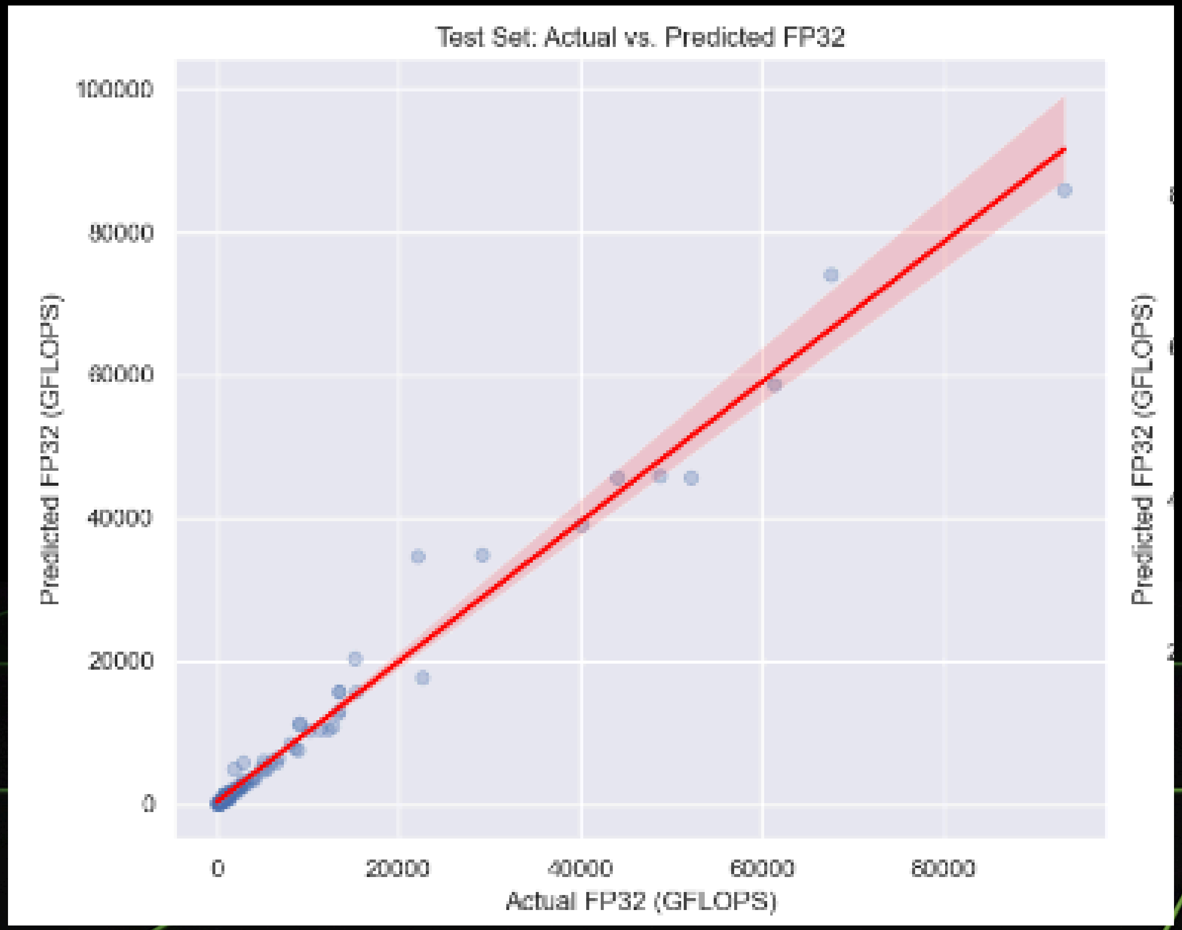
Random Forrest Regression:	XGBoost:	Gradient Boosting Regression:
Train RMSE: 1237.7696	Train RMSE: 433.2180	Train RMSE: 287.0675
Train R^2: 0.9896	Train R^2: 0.9987	Train R^2: 0.9994
Test RMSE: 3009.0642	Test RMSE: 2130.4593	Test RMSE: 3077.9636
Test R^2: 0.9623	Test R^2: 0.98110	Test R^2: 0.9605

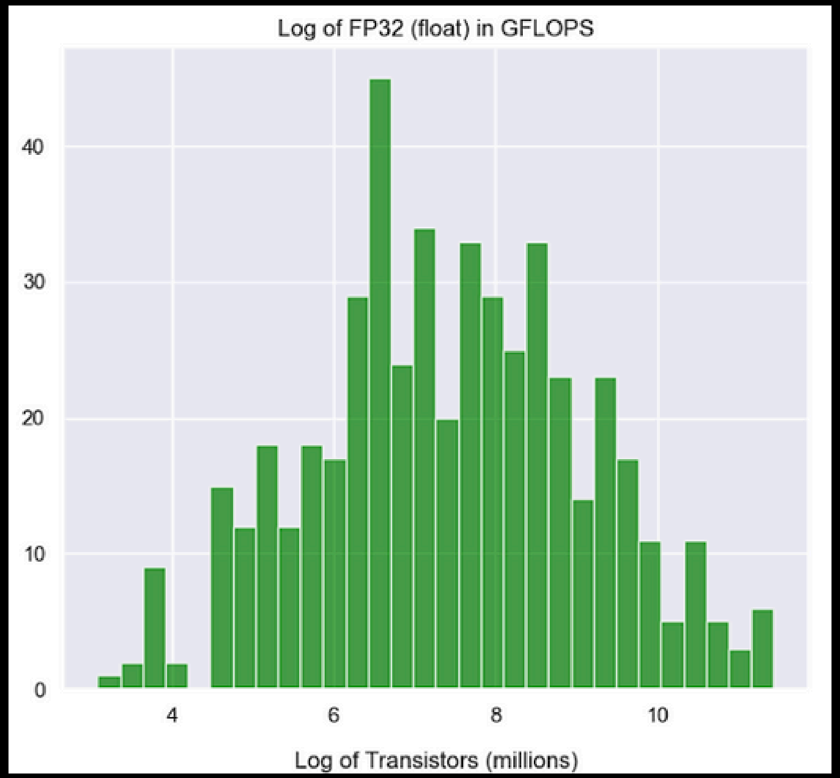
What Does It Mean?

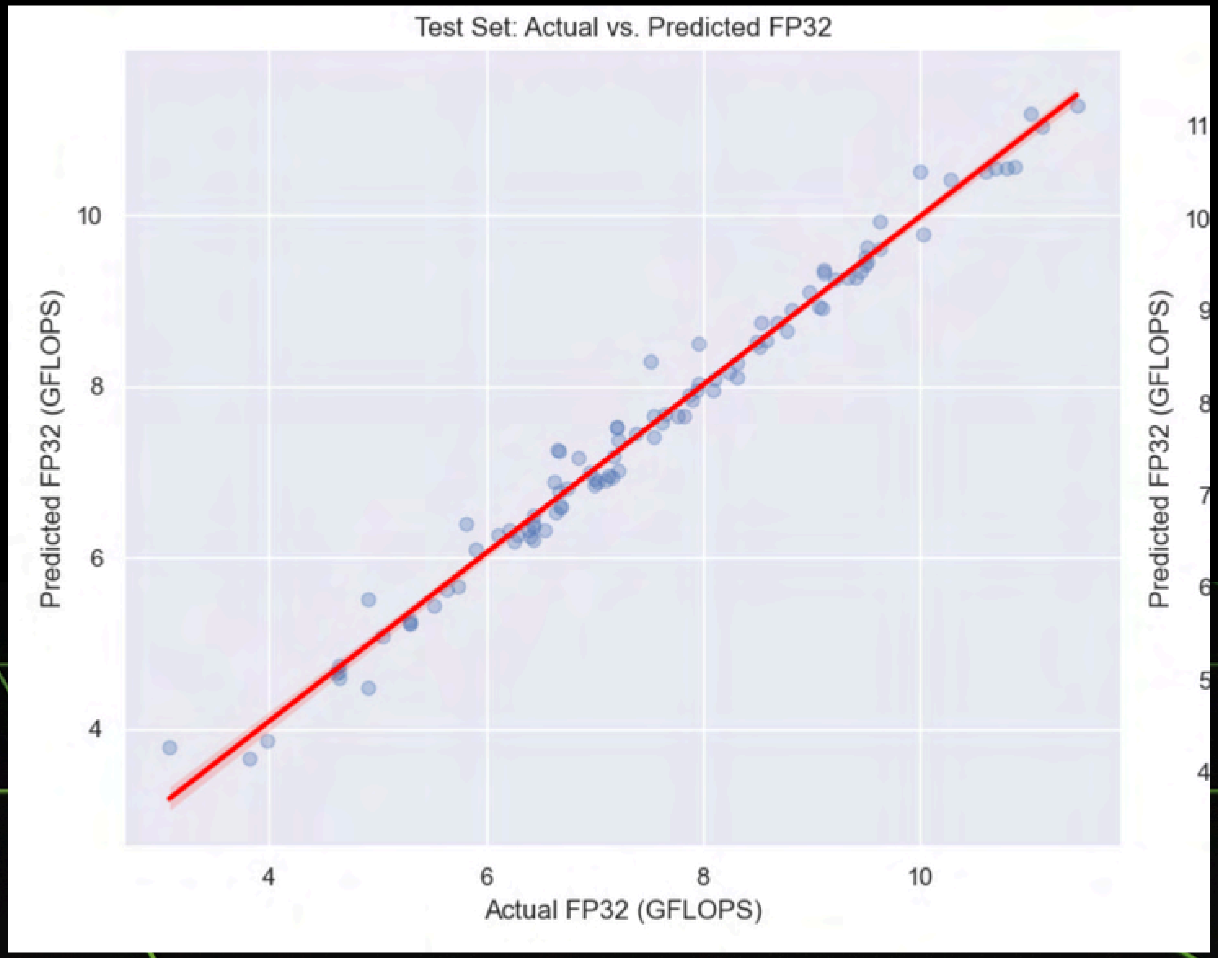
- Random Forest Regression shows a high degree of overfitting
- XGBoost offers an improvement in generalisation over Random Forest
- Gradient Boosting's RMSE suggests excellent performance on the training data but exhibits potential overfitting
- In summary, XGBoost is the most balanced choice with a mix of high accuracy and generalization capabilities

SKEW CORRECTION









FEATURE IMPORTANCE

Random Forest Regression:
First Model (max depth 4):

Transistors (millions): 0.9029
GPU clock (MHz): 0.0550
Memory Size (MB): 0.0198
Bus Width (bits): 0.0077
Memory clock (MHz): 0.0058
Density (M/mm²): 0.0054
Max. TDP (W): 0.0019
Die Size (mm²): 0.0007
Process Size (nm): 0.0004
L2 Cache (KB): 0.0004
Gen Year: 0.0001

XGBoost:

Transistors (millions): 0.9104
GPU clock (MHz): 0.0336
Memory Size (MB): 0.0098
Memory Type_GDDR6X: 0.0066
Max. TDP (W): 0.0051
Density (M/mm²): 0.0044
Memory clock (MHz): 0.0039
L2 Cache (KB): 0.0037
Architecture_RDNA 3.0: 0.0035
Die Size (mm²): 0.0029
Bus Width (bits): 0.0025

Gradient Boosting Regression:

Transistors (millions): 0.9252
GPU clock (MHz): 0.0319
Memory Size (MB): 0.0212
Bus Width (bits): 0.0140
Memory clock (MHz): 0.0041
Memory Type_GDDR6X: 0.0006
Density (M/mm²): 0.0005
Architecture_RDNA 3.0: 0.0005
L2 Cache (KB): 0.0004
Max. TDP (W): 0.0004
Architecture_Ampere: 0.0003

MULTI-OUTPUT REGRESSION

Variables related to response variables:

- FP32 (float) in GFLOPS
- Pixel Rate (GPixel/s)
- Texture Rate (GTexel/s)

Feature Importance:

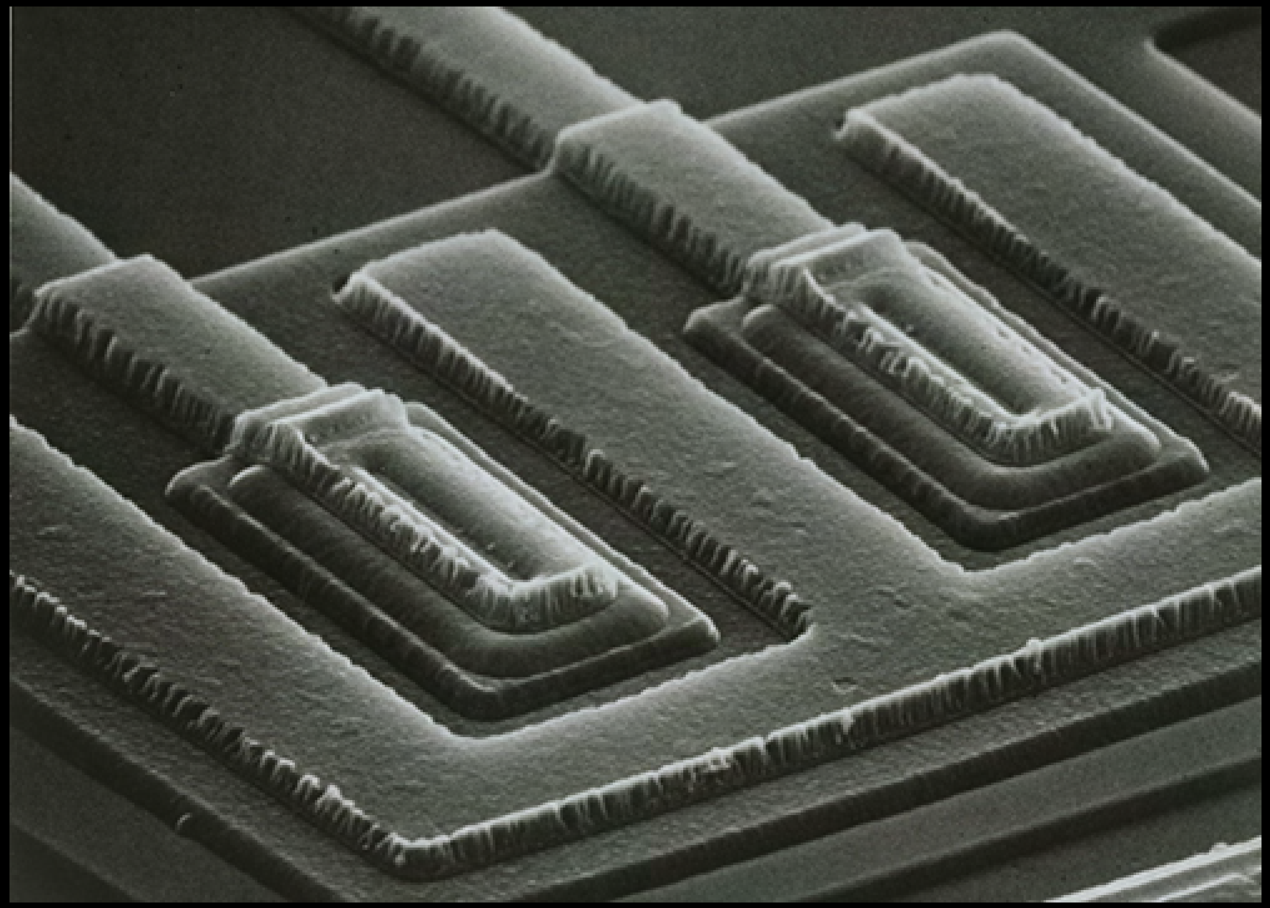
Transistors (millions): 0.7439350
Memory Size (MB): 0.14394547
Memory Type_HBM2: 0.0155372
GPU clock (MHz): 0.01534710
Manufacturer_AMD: 0.0113193
Memory clock (MHz): 0.00983113
Foundry_Samsung: 0.00760316
Architecture_Ampere: 0.0062706
Density (M/mm²): 0.0062431
Bus Width (bits): 0.005865

DATA DRIVEN INSIGHTS

We noticed that

- Transistors (millions) and GPU Clock (MHz)
- Memory configuration
- Architecture

had high feature importance across all models.



OUTCOME & CONCLUSION

- A model that allows manufacturer to predict the FP32 of a chip based on its proposed features.

- GPU manufacturers must aim to improve transistor density and clock speeds for best performance uplift

- Emphasis should be placed on faster, larger capacity and higher bandwidth memory. Faster memory means the GPU is able to access information faster, leading to greater performance uplift

- More efficient and advanced architectures allow for higher throughput while providing powerful and exciting feature sets.

- So continuous investment in research and development to innovate GPU will lead to better market positioning and product performance.

THANK YOU!