EP2420 *Project 1 - Week 1*

Federico Giarrè

November 2, 2023

1 Data Exploration

1.1 Composition of the datasets

The datasets X and Y composed of 3600 observations each. In Table 1 a statistical overview of the composition of the dataset is given.

Feature	Mean	Std Deviation	Max	Min	20th %-tile	50th %-tile	90th %-tile
runq-sz	5.27e+01	4.74e+01	1.99e+02	3.00e+00	1.10e+01	3.30e+01	1.26e+02
%%memused	2.62e+01	6.25e+00	3.90e+01	1.50e+01	2.29e+01	2.58e+01	3.59e+01
proc/s	6.18e+00	9.06e+00	5.00e+01	0.00e+00	0.00e+00	0.00e+00	1.90e+01
cswch/s	4.31e+04	2.47e+04	8.50e+04	8.38e+03	1.88e+04	3.88e+04	7.33e+04
all_%%usr	7.31e+01	2.77e+01	9.86e+01	1.75e+01	4.50e+01	9.35e+01	9.81e+01
ldavg-1	6.04e+01	5.04e+01	1.87e+02	4.60e+00	1.34e+01	4.40e+01	1.39e+02
totsck	4.55e+02	1.81e+02	9.58e+02	2.69e+02	3.11e+02	3.66e+02	7.53e+02
pgfree/s	7.39e+04	2.87e+04	4.43e+05	1.67e+04	5.61e+04	7.15e+04	1.13e+05
plist-sz	8.02e+02	3.46e+02	1.77e+03	4.52e+02	5.25e+02	6.32e+02	1.37e+03
file-nr	2.64e+03	1.97e+02	3.31e+03	2.35e+03	2.50e+03	2.59e+03	2.93e+03
idel/s	4.56e+01	2.56e+02	9.01e+03	1.00e+00	1.10e+01	2.10e+01	6.10e+01
tps	6.25e+00	1.27e+01	9.00e+01	0.00e+00	0.00e+00	0.00e+00	1.80e+01
DispFrames	2.03e+01	4.99e+00	3.04e+01	0.00e+00	1.48e+01	2.34e+01	2.50e+01

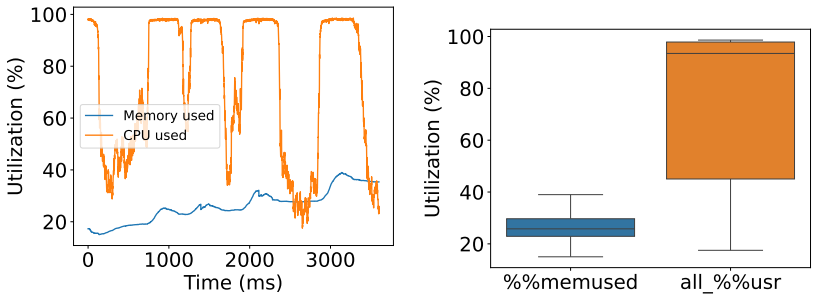
Table 1: Statistics of the composition of the datasets

1.2 Subsets of dataset X

a) The amount of observations where the CPU usage is below 50%, and Memory usage is below 25% is **4.49e+02**;

b) The average amount of used sockets when the number of context switches per second is below 50.000 is **3.21e+02**.

1.3 Memory and CPU analysis

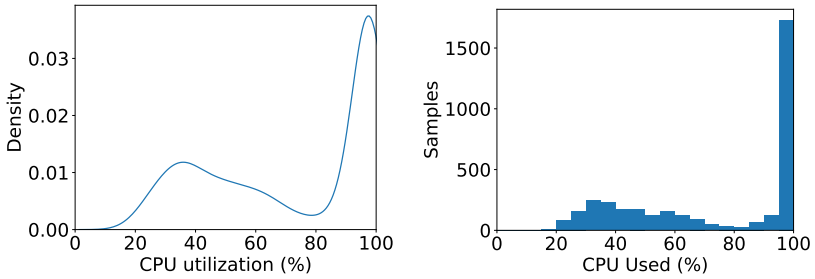


(a) CPU and Memory behaviors with respect to time

(b) Boxplot of CPU and Memory usage values

Figure 1: Time series and Box plot of both Memory and CPU usage

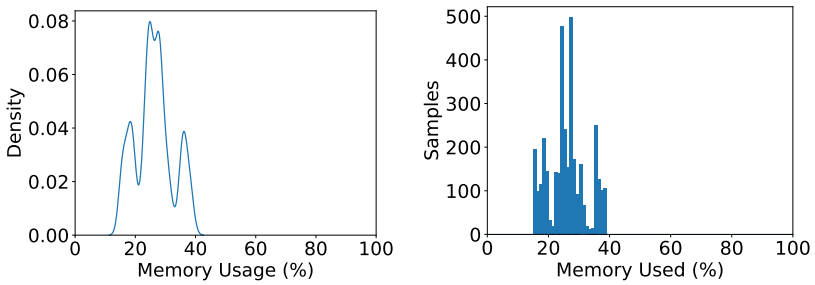
In figures 1(a) and 1(b) it is possible to notice how the two metrics of Memory and CPU usage differ in both values, trends, and distribution. For instance, as it is possible to see in Fig. 1(a) Memory usage increases over time while CPU tends to peak and drop frequently. This is also reflected in Fig. 1(b), where Memory is concentrated between 20% and 30% of utilization while the median of the CPU utilization is 93%. As a consequence of the behavior of the CPU noticeable in 1(a), the values below the median are very sparse with the lower quartile reaching 45%.



(a) CPU usage density plot

(b) CPU usage Histogram

Figure 2: Density and Histogram for the distribution of values of CPU usage



(a) Memory usage density plot

(b) Memory usage histogram

Figure 3: Density and Histogram for the distribution of values of Memory usage

In Figure 2, a density plot (Fig. 2(a)) and a histogram (Fig. 2(b)) are plotted in order to study the distribution of values of the CPU usage present in the dataset. The same plots are have also been created for the Memory usage values (Fig. 6). The CPU histogram and density plot highlight the conclusions given on the boxplot of Fig. 1(b), the majority of observations are on the peak, with the rest of the data spreading along the other values. While the memory's histogram and density plot show some smoother distribution of values, there are still some gaps in the memory usage in between the more common values.

1.4 Answer to the quiz question

As seen in the previous analysis, the values for the features can be sparse, dense, have peaks or be more uniform, have gaps or not. A linear regression model trained on the data of this dataset will be accurate when giving a prediction on features it has already seen (e.g. high CPU load and low Memory usage) but will have no clue on how to interpret features such as high memory usage or low CPU consumption thus giving prediction that are not accurate.

2 Estimating Service Metrics from Device Statistics using Linear Regression

A Linear regression model was created in order to give predictions about the VoD performance given the VoD server statistics. In order to train such model the dataset was split into Training set (70% of the initial dataset) and a Test set (30% of the initial dataset).

2.1 Evaluate the Accuracy of Service Metric Estimation

2.1.1 Model training

After training the model, the resulting coefficients are: (-4.3e-04, -3.3e-02, 1.4e-02, 1.0e-02, -1.8e-04, 1.3e-01, -1.5e-02, 2.3e-02, -9.6e-06, -1.5e-02, 3.7e-03, -5.2e-05, -2.5e-03)

2.1.2 Accuracy of model M

The values for the Normalized Mean Absolute Error (NMAE) for both the trained model and a naïve prediction (mean value of the labels in the training set) can be found in Table 2.

Method	NMAE
LR Model	9.81e-02
Naïve	2.27e-01

Table 2: NMAE values for the linear regression model and naïve predictions

2.1.3 Prediction

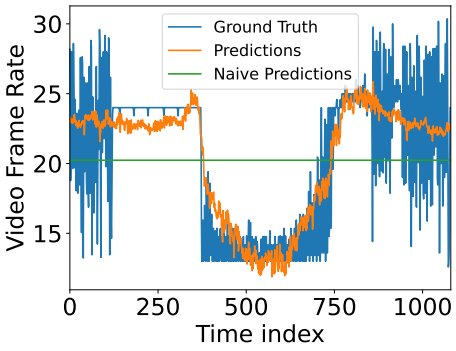
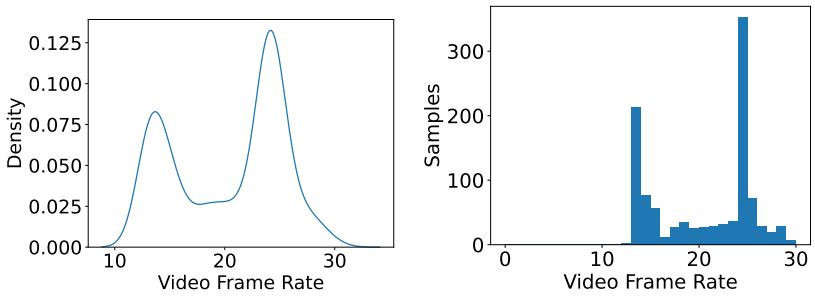


Figure 4: Comparison of measured values, model prediction and naïve prediction on the testset

2.1.4 Testset composition

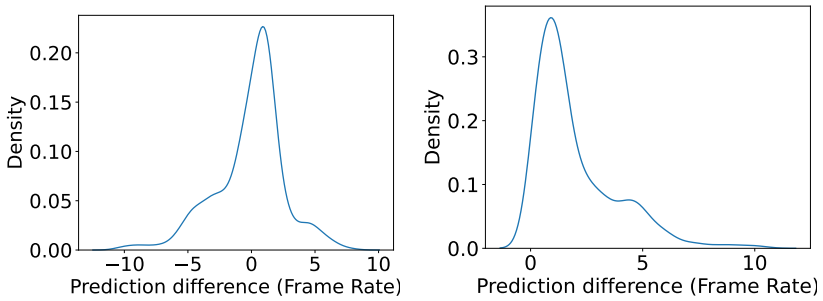


(a) Testset density plot

(b) Testset histogram

Figure 5: Density and Histogram for the distribution of values in the Testset

2.1.5 Measurement / Prediction difference



(a) Prediction difference

(b) Absolute prediction difference

Figure 6: Distance between model prediction and ground truth

2.1.6 Conclusions

As expected by the NMAE values in Table 2, Fig. 4 shows that the Naïve prediction can not grasp the trend of the real data. In contrast, the model predictions manage to follow the behavior of the real measurements better. However, the mean Frame Rate of the training set is around 20 FPS (as visible from the predictions of the naïve method), while the test set contains few sample values around that one as seen in Fig. 5(a) and 5(b). This results in the model being accurate when predicting the trend, but not the exact value (as seen by the model not catching peaks and lows of the test set). The prediction is not always value-accurate with 30% of the predictions being 1 FPS away from the measurement, and other predictions reaching up to 10 FPS of difference from the ground truth.