EP2420 - Project 1: Task 1, Task 2.1

André Silva

October 31, 2020

Task I

1. The dataset *VoD flashcrowd* provides us with 36633 samples of 1670 features, and 9 different types of targets. In Table 1, the statistics for the chosen features, as well as the target *DispFrames*, which represents the video frame rate, are displayed.

Feature name	Mean	Std Dev	Maximum	Minimum	25th percentile	90th percentile
3_cpu16_.sys	3.55	1.95	1.50×10^1	0.00	0.00	0.00
23_RxPacktes	3.15×10^2	4.12×10^2	2.37×10^3	0.00	3.70×10^1	4.70×10^1
3_frmpg.s	-1.15×10^1	2.07×10^4	1.08×10^5	-7.49×10^4	-5.31×10^4	-4.55×10^4
0_cpu18_.idle	9.97×10^1	5.98×10^{-1}	1.00×10^2	9.50×10^1	9.70×10^1	9.80×10^1
2_cpu17_.usr	6.03×10^1	1.82×10^1	9.20×10^1	0.00	8.00	1.34×10^1
4_cpu15_.iowait	4.65×10^{-3}	7.21×10^{-2}	3.03	0.00	0.00	0.00
4_cpu5_.sys	2.90×10^1	1.08×10^1	7.40×10^1	0.00	6.06	9.18
3_cswch.s	7.30×10^4	1.79×10^4	1.07×10^5	7.82×10^3	2.19×10^4	2.65×10^4
38_TxBytes	2.79×10^6	3.18×10^6	1.96×10^7	0.00	3.47×10^5	4.63×10^5
2_dev8.0_avgrq.sz	2.21×10^1	5.78×10^1	1.02×10^3	0.00	0.00	0.00
DispFrames	2.20×10^1	4.32	2.50×10^1	0.00	3.00	8.00

Table 1: Statistics for chosen features and target

The following list gives a short description of these features. This information was retrieved from linux manual pages for the command `sar` [1].

- *3_cpu16.sys* - Percentage of CPU utilization that occurred while executing at the system level (kernel).
- *23_RxPackets* - Total number of packets received per second.
- *3_frmpg.s* - Number of memory pages freed by the system per second.
- *0_cpu18_.idle* - Percentage of time that the CPU or CPUs were idle and the system did not have an outstanding disk I/O request.
- *2_cpu17_.usr* - Percentage of CPU utilization that occurred while executing at the user level (application).
- *4_cpu15_.iowait* - Percentage of time that the CPU was idle during which the system had an outstanding disk I/O request.
- *4_cpu5_.sys* - Percentage of CPU utilization that occurred while executing at the system level (kernel).
- *3_cswch.s* - Total number of context switches per second.
- *38_TxBytes* - Total number of bytes transmitted per second.
- *2_dev8.0_avgrq.sz* - The average size (in sectors) of the requests that were issued to the device.

Task II - 2.1

4. Table 2 provides the calculated *Normalized Mean Absolute Error* (NMAE) for each of the regressors. The parameters utilized for the random forest regression and the neural network regression are the default parameters of, respectively, `RandomForestRegressor` [2] and `MLPRegressor` [3], unless specified in the table.

Regressor	NMAE
`LinearRegression`	0.318
`RandomForestRegressor(n_estimators=10)`	0.087
`MLPRegressor(max_iter=1000, activation='logistic', hidden_layers=(10,10))`	0.144

Table 2: *Normalized Mean Absolute Error* for each regressor tested

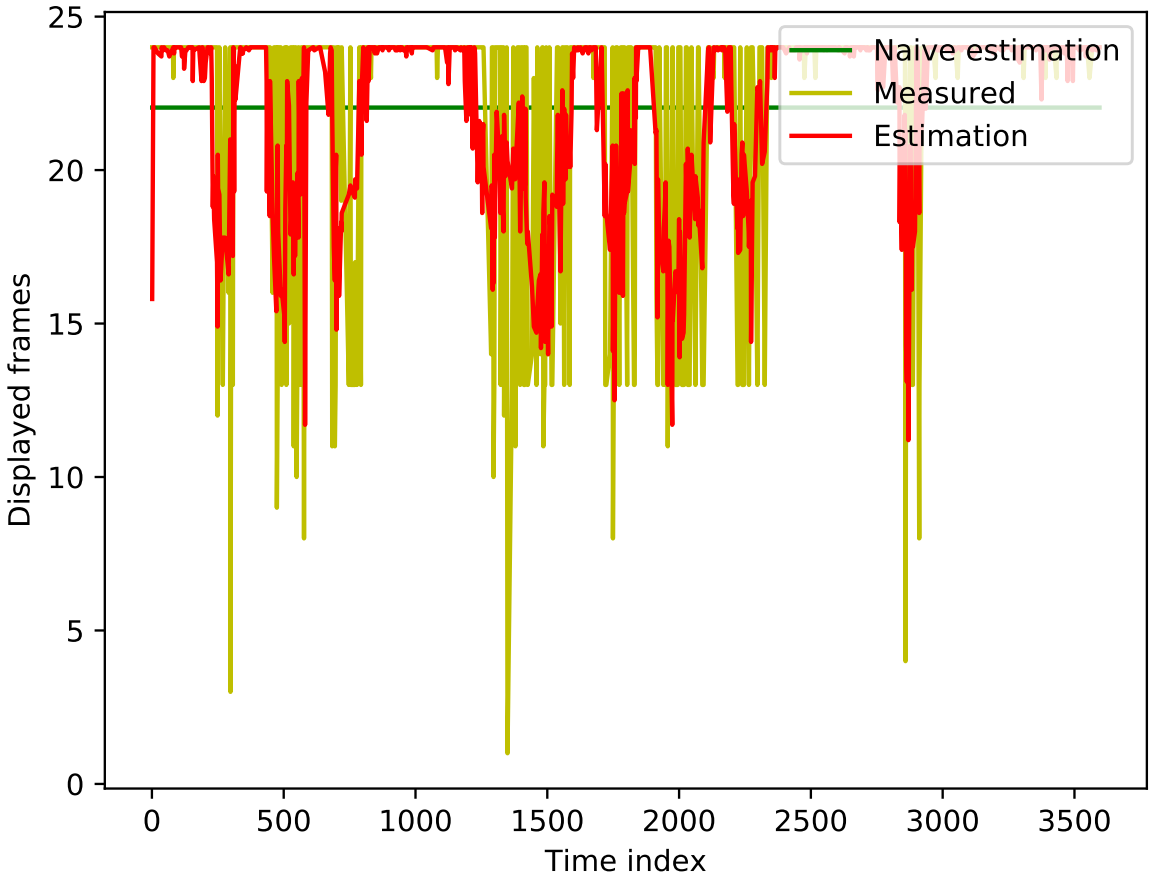

Figure 1: Time series plot for *VoD flashcrowd* using `RnadomForestRegressor`

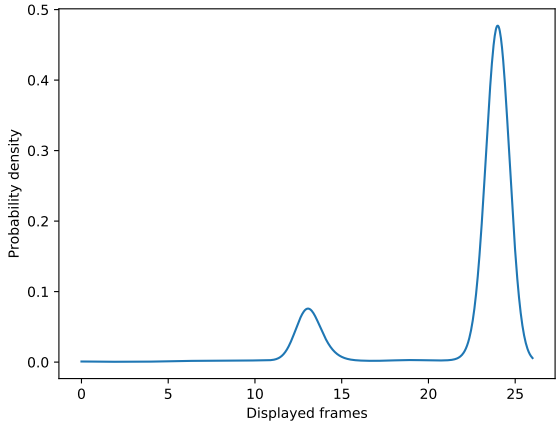

Figure 2: Density plot of the target values in the test set

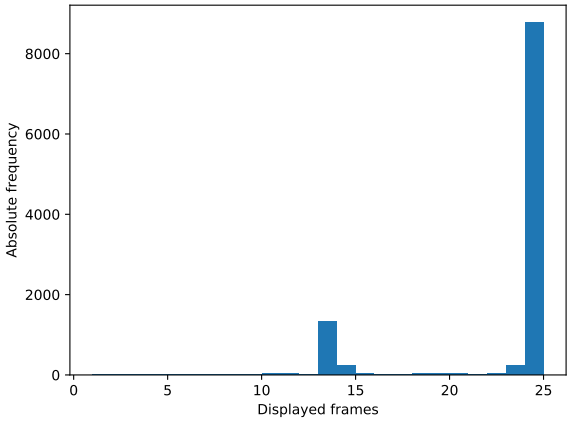

Figure 3: Histogram of the target values in the test set

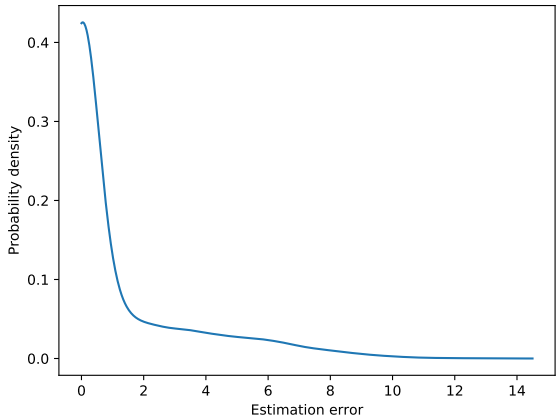

Figure 4: Density plot of the estimation errors $y_i - \hat{y}_i$ in the test set

9. Of the three regression techniques utilized, linear regression was the least expensive in computational power, but also the one which provided the worst accuracy.

   Neural network regression was more expensive when compared to random forest regression, but provided worse accuracy.

   Random forest regression was better both in terms of accuracy and cost-performance trade-off when compared to the other two models.

References

[1] *`sar` command manual page*. `https://linux.die.net/man/1/sar`. accessed: 2020-10-31.

[2] *sklearn.ensemble.randomforestregressor*. `https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.randomforestregressor.html#sklearn.ensemble.randomforestregressor`. accessed: 2020-10-31.

[3] *sklearn.neural_network.MLPRegressor*. `https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPRegressor.html#sklearn.neural_network.MLPRegressor`. accessed: 2020-10-31.