Class_ID:8-Kernel run time prediction

I have selected SGEMM GPU kernel performance dataset which was released by UCI machine learning repository (https://archive.ics.uci.edu/ml/datasets/SGEMM+GPU+kernel+performance#) this data set was released last month where they measured run time of kernel using various combination of kernel features which I will elaborate while explaining attributes.

Datasets have 241600 instances and 18 attributes
18 attributes includes 10 ordinal 4 binary and last 4 attributes are run variable.
So basically every combination of 14 attributes were run 4 times and there result were recorded.

Analysis:
Method1:
Main task would be to see the linear relationship with all 14 attributes and 4 different run times. Linear regression would be best method for this task.
Second want to find irrelevant predictors from attributes which don't have impact on run time of the kernel.

For this shrinkage method can be applied such as PCA ,PCS LASSO regression. Even ridge regression can be applied for this task.
Shrinkage and regularization method would help us know which attribute does not have any weightage on output. This will help manufacturer to focus on that particular attributes for reducing run time.

Clustering: Clustering can be applied for similar attributes and can be seen that run time is same for all those attributes or run times differ within clusters.

Difference in run time for instance run1-run4 and we can see if model is able to capture the variance of difference between run1 and run4. Secondly we can take average of all 4 runs and than feed into linear model .

Cross Validation and LOOCV or validation set and bootstrap can be applied to reduce the error studied by model and increase the accuracy to balance bias variance trade off.

If uncleaned data is found than value can be using that column in the dataset but this can be risky as majority of attributes are ordinal which can have large attributes on till now data is cleaned.

Xgboost can be used to improve accuracy as it performs feature engineering and tries to capture all the variance left by norml linear regression.