DSCI799 Data Profiling Report

Evaluating EXAMM on Coal Powerplant Data

Introduction

The dataset I will use for this project is called the “Coal Burner Dataset”. It has been used as a benchmark for EXAMM in many published papers and I intend to continue that tradition. The dataset is a collection of minute - by - minute readings taken from 12 sensors in a coal powerplant burner over 120 days. The sensors used in this dataset are listed in Table 1. Each 10 - day period is stored in an individual csv file, giving us a total of 12 csv files with 14,401 records each (172,812 total). Typically the first 10 files have been used for training and the last 2 have been used for testing.

Sensors	
1. Conditioner Inlet Temperature	7. Secondary Air Flow
2. Conditioner Outlet Temperature	8. Secondary Air Split
3. Coal Feeder Rate	9. Tertiary Air Split
4. Primary Air Flow	10. Total Combined Air Flow
5. Primary Air Split	11. Supplementary Fuel Flow
6. System Secondary Air Flow Total	12. Main Flame Intensity

Table 1: Sensors in Coal Powerplant Dataset

Typically, readings from sensors 1 to 11 are used as predictor variables and the reading from sensor 12 (main flame intensity) is the response. This is not set in stone however and we can set any one of the sensors to be the response. That being said, I will stick to using the main flame intensity as the response variable.

Basic Analysis and Summary Statistics

The dataset has n = 14401*12 = 172812 rows and p + 1 = 12 columns, where p = 11 is the number of predictor variables. All of the variables are numeric and have already been normalized to be between zero and one. There are no nulls in the data and the typical train:test ratio is 5:1, although this can be changed if desired. Table 2 presents summary statics for all 12 sensors. Please note that values such as min, max, and range are not included since the readings have already been normalized, and the min, max, and range are 0, 1, and 1 respectively for all 12 sensors. Furthermore, to help visualize the data, the time plots of each variable are presented in Figure 1. These show the evolution of each variable over the 120 days and give us a rough idea of the distribution of each variable and their correlations, although these aspects will be studied individually.

Sensor	Median	Mean	Mean 95% CI	Variance
1	0.46923	0.48521	0.00082	0.03009
2	0.34809	0.34323	0.00031	0.00431
3	0.91289	0.86973	0.00091	0.03742
4	0.72856	0.68855	0.00067	0.02019
5	0.98739	0.97610	0.00028	0.00354
6	0.76860	0.71826	0.00074	0.02446
7	0.65095	0.60648	0.00063	0.01758
8	0.01467	0.02323	0.00025	0.00276
9	0.00848	0.03362	0.00049	0.01078
10	0.71471	0.67114	0.00069	0.02116
11	0.00326	0.00323	0.00006	0.00014
12	0.82401	0.79767	0.00061	0.01672

Table 2: Summary Statistics

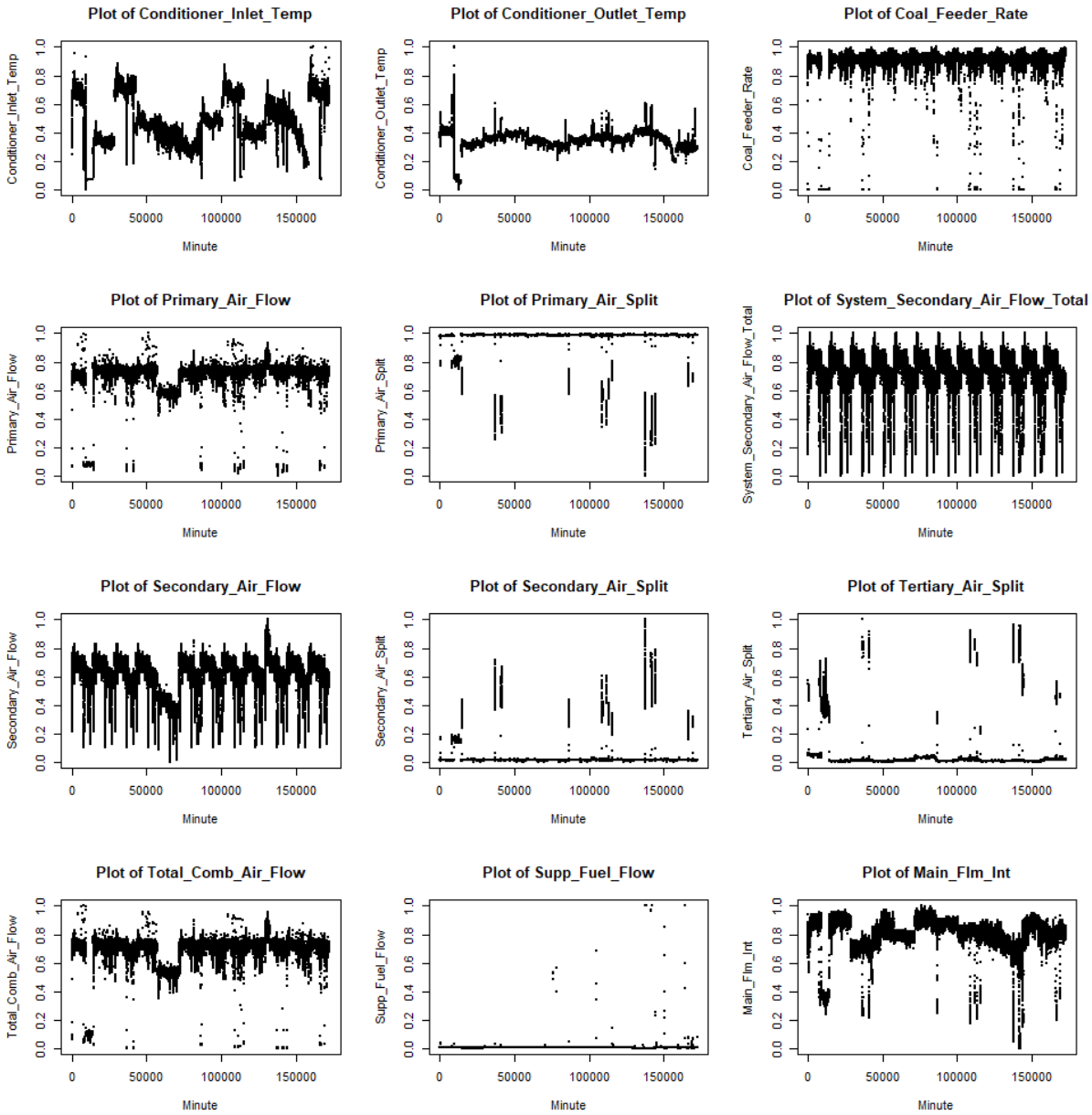

Figure 1: Plots of all Variables in Dataset

Distribution of Variables

The histograms shown in Figure 2 help visualize the frequency distribution of each variable. We see that some variables are more widely distributed while others almost always take on the same value, which mirrors what we saw from the time plots in Figure 1.

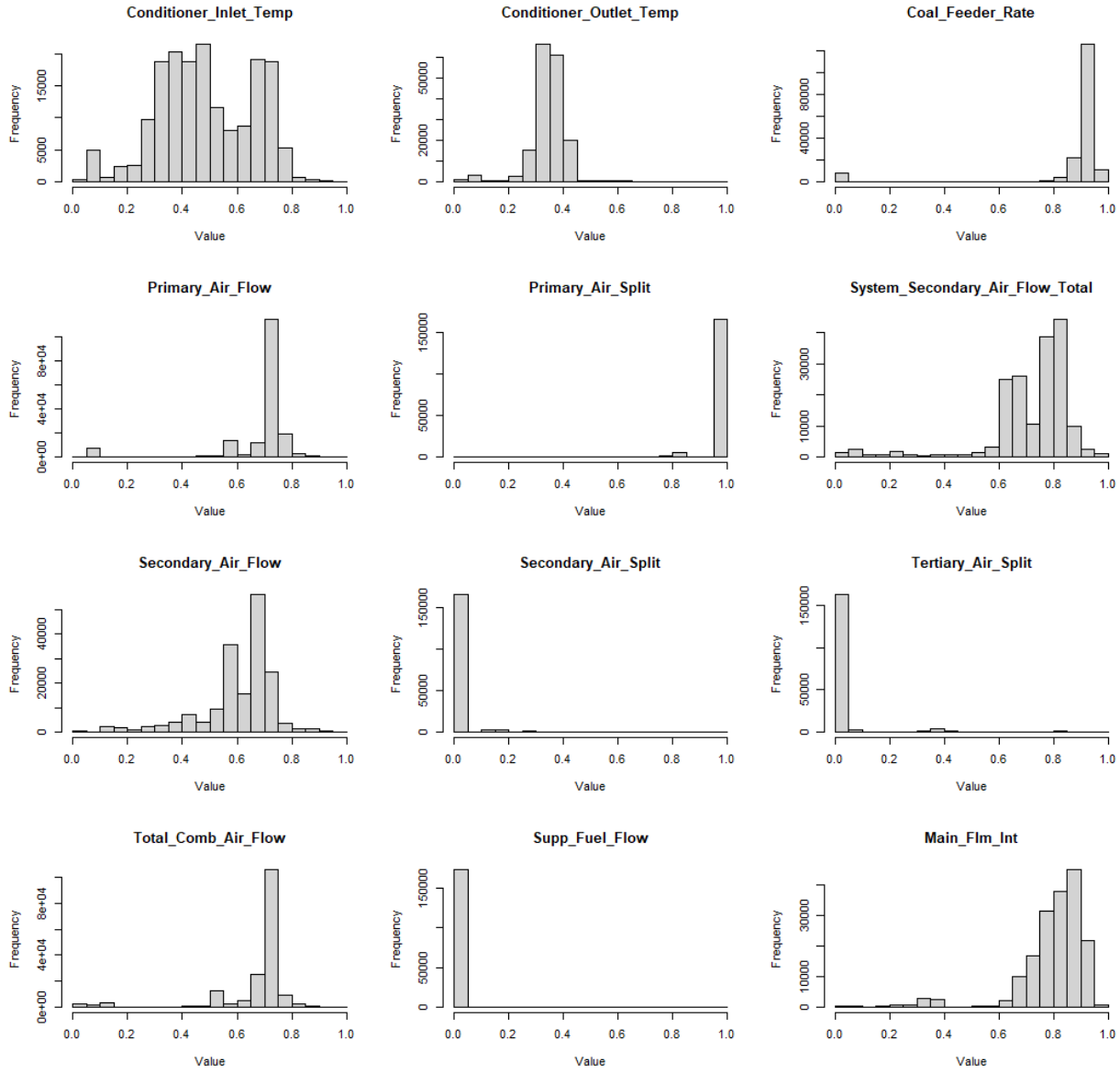

Figure 2: Histograms of all Variables in Dataset

Correlations

We can determine variable correlations by generating a correlation matrix, as shown in Figure 3. It is important to keep that in the case of time series data, the correlation matrix does not provide all that much info. Correlations are calculated based on the values of two variables at the same index/timestep, whereas in time series data the value of the response depends not on the predictor values at the same timestep, but on time – lagged values of the predictors and the response itself. One potential use for the correlation matrix is to determine which predictors are redundant and may be eliminated to reduce complexity. That being said, I will not be considering dimensionality reduction to ensure that I use the same data as other implementations of EXAMM.

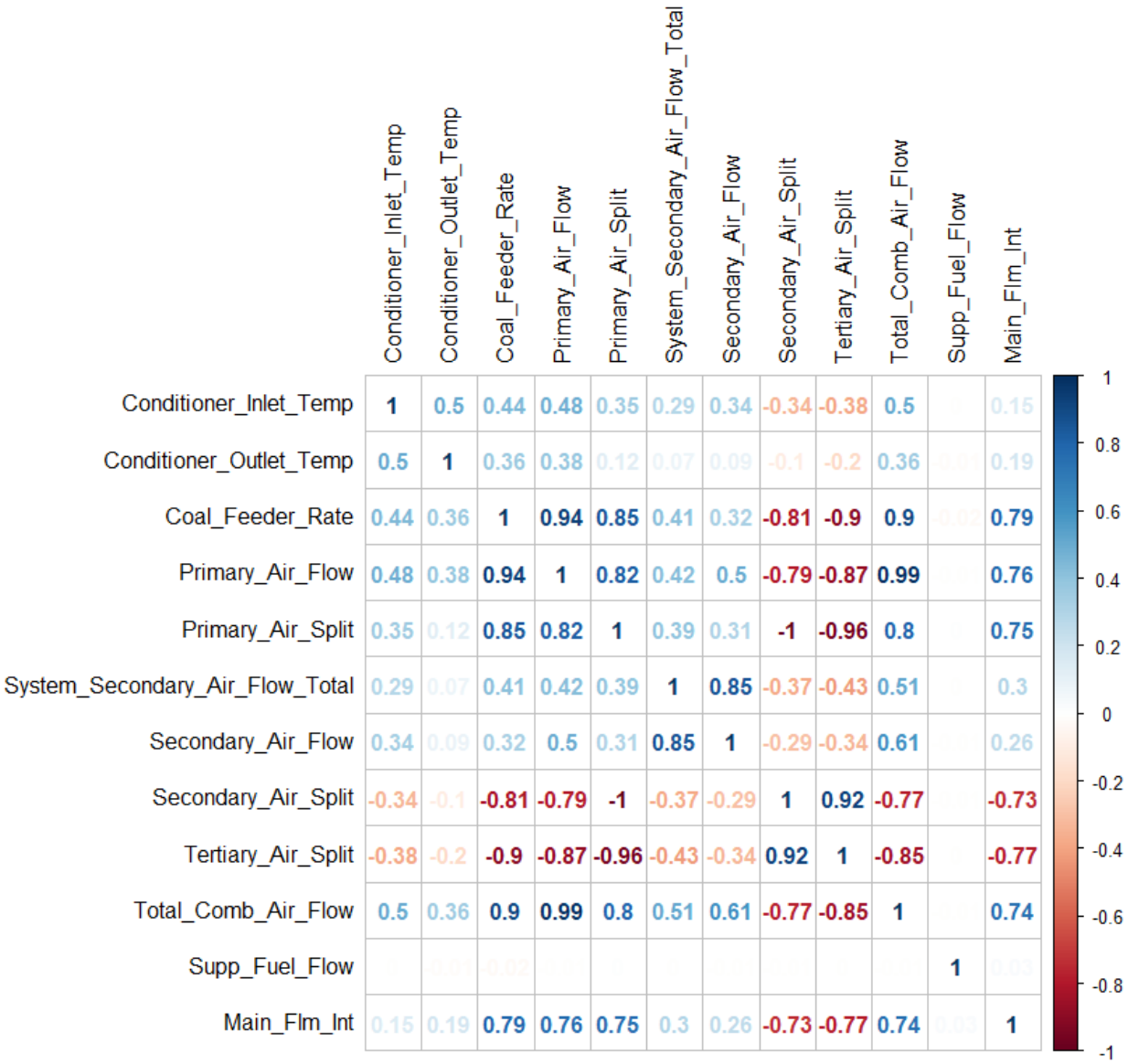

Figure 3:Correlation Matrix of Coal Powerplant Data

Concluding Remarks

Overall I think this is a very friendly dataset, as it is already normalized and has no null values. I suppose this explains why it was chosen as one of the benchmark datasets for EXAMM. It is relatively low dimensional and there is a lot of data available to train models on, although it is but a fraction of the million rows suggestion given in week 2. That being said, I'm confident that this dataset will work for this project. It has been used to profile and test EXAMM numerous times in the past and I don't foresee any issues. As already mentioned, the data is clean and normalized, so as far as I can tell there is nothing to correct. One point of interest is the amount of correlated variables, such as the extremely high correlation between the primary air flow and total combined air flow predictors. If I was doing a classical data science project I would consider eliminating some of the correlated variables in order to reduce complexity. However, because of the nature of my project, I will use the full dataset to ensure a fair comparison.