

STEAM VOLUME PREDICTION FOR INDUSTRIAL BOILERS

Bolin He, Xiaoyuan Jiang, Yidong Li

University of California San Diego, La Jolla, CA 92093

ABSTRACT

The main purpose of this project is to investigate the efficiency of a boiler under various parameters by predicting the steam volume. The efficiency of boilers, which is also called total thermal energy, is complicated and difficult to be determined, because boiler loss is caused by various factors including combustion air feed rate, radiation loss, and ambient temperature [1]. One way to evaluate the efficiency is to predict the steam volume under different parameters with machine learning [2]. We tried to clean the data and used several machine learning models to predict the efficiency.

Index Terms—steam volume, Box-Cox transformation, Ridge Regression, One-Class SVM, Linear Regression, Gradient Boosting Regression, MLP

1. INTRODUCTION

This project is based on an industrial model that investigates the efficiency and steam volume of boilers for a thermal power plant, along with the prediction of those features under various of circumstances. Boilers are essential devices in diverse aspects of industrial procedures such as electric power networks and transportation engineering. Thus they are not only the core equipment to convert the energy but also important steam generators in certain scenarios. In this project, the scope is limited to a boiler for a coal-fired power plant. The basic principle of a thermoelectricity power generation is: firstly, burning fuel to heats water to generate steam. Then the steam pressure pushes the steam turbine to rotate, and finally the steam turbine drives the generator to rotate to generate electrical energy [3].

All the parameters such as combustion air feed rate [1], radiation loss and the ambient temperature in this process along with the steam volume itself are mutually influencing each other. These interrelationships make it not reasonable to separate the focused outcomes out of the other parameters. Thus, the existing values of all those mentioned parameters as well as the steam volumes are the input of our algorithm, which are discussed in details in the dataset section as the 38 features, while the outputs are their presentation according to the interrelationship behaviors. These outputs were predicted by applying and comparing different models including linear regression, gradient boosting regression, MLP and so on.

2. RELATED WORK

The project was limited under the scope of thermal power plant after the literature reviews. According to the gathered materials, the boilers for thermal power generators are the most reviewed aspects among the topic. It also includes the most sufficient data and complete algorithms. All these make it the optimal resource for the learning materials. By and large, the methodology to investigate the features of a certain boiler can be classified as manual experimental works based on variable controlling methods and presentation based on deep-learning. Those relatively old materials are mostly based on complicated experimental methods. [4] and [5] are both typical researches on this topic by manual experiment.

The drawbacks for variable controlling and manual experiments are obvious. First, the efficiency of experiment is seriously limited by the time wasted on debugging the system and shifting the parameters manually. Also the outcomes cannot be accurate enough since most of the parameters are interacted with each other and hard to be changed independently as mentioned. In the researches in this century, however, various detectors and sensors give the possibility to determine the influence of all the parameters simultaneously. This, on the one hand, significantly increased the performance of optimal choice of a boiler can make, on the other hand, dramatically enlarged the data to be processed so as to predict the steam volume, which was far beyond manual experiment's achieve. All these factors cause the application of Deep Learning in this topic. Among these works, different sorts of regressions are applied in the majority of scenarios such as [6] and [2]. Nevertheless we also tried other models like random forest.

3. DATASET AND FEATURES

The data we use is some masking data collected by the boiler sensors with the collection frequency on the minute level. The dimension of training set is (2888, 39), which means we have 2888 training sample with 39 channels. 38 of the channels are features starting from V0 to V37, with the 39th channel as target, the actual amount of steam volume. Similarly, the dimension of test set is (1925, 38) without the 39th target channel. In this case, we only use test set during data preprocessing. For model training, we split the training set into a sub training set and a sub validation set.

3.1. Features Visualization

For the dataset, the first thing we do is some data preprocessing. We try to find out the data distribution from V0 to V37. The Fig. 1 shows the distribution (i.e. the frequency of each values) of V0, V1, V5, V11. For V0 and V1, the training and test data almost have the same distribution while V5 and V11 show the different distribution of data.

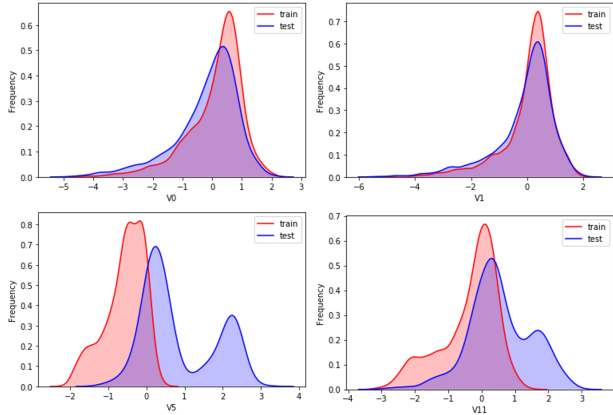


Fig. 1. Data distribution of V0, V1, V5, V11

Through this visualization, we realize that data preprocessing is a significant and necessary procedure. We find that some branches of features are relatively irrelevant since their distribution of training set and test set are not close to each other. This kind of features will cause overfitting which will make the model work well in the training set but have a bad result in the test set. In the dataset we use, features of V5, V9, V11, V14, V17, V21, V22, V28 have the distribution difference between training set and test set. Finally, we decide to eliminate these features.

3.2. Correlation coefficient

Fig. 2 shows the correlation coefficient matrix between every two features among all the rest 30 feature channels. For both the red and blue blocks, the darker the blocks, the more relevant the two features, and vice versa. It is important to illustrate the correlation among them. Poor correlation indicates that two features are irrelevant enough, which may decrease model accuracy and increase running time. We decide to drop the redundant features V25, V26, V32, V33, V34 with an absolute threshold of 0.1.

3.3. Normalization

For the normalization, we normalize the data into 0 to 1 range. The goal of normalization is to change the values of numeric columns in the dataset to a common scale, without distorting differences in the ranges of values. Normalization makes

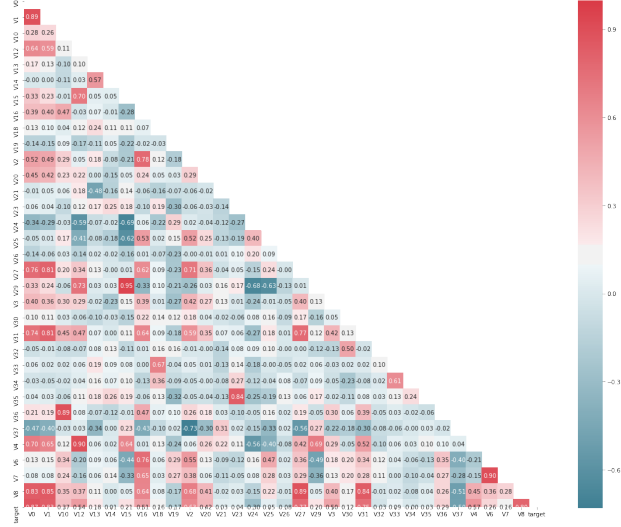


Fig. 2. Correlation coefficient matrix

sure that all of the data look and read the same way across all records.

3.4. Box-Cox transformation

A Box-Cox transformation is a way to transform non-normal dependent variables into a normal shape. Normality is an important assumption for many statistical techniques. It follows the transformation formula as:

$$y(\lambda) = \begin{cases} \frac{y^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ \ln y, & \lambda = 0 \end{cases} \quad (1)$$

After Box-Cox transformation, Fig. 3 shows the distribution of the first five features of data before and after the transformation. A positive result is found as those distributions are in better normal shapes after Box-Cox transformation.

3.5. Ridge and One-Class SVM

Finally, we try to eliminate the outliers through two steps: Ridge Regression and One-Class SVM.

Ridge Regression, also known as Tikhonov regularization, however, may achieve the same goal by modifying linear least squares loss function and a L2-norm regularization, and thus contribute to the trade-off between bias and variance.

$$\sum_{i=1}^n (y_i - \sum_{j=0}^p w_j x_{ij})^2 + \lambda \sum_{j=0}^p w_j^2 \quad (2)$$

$$\begin{aligned} & \|Xw - y\|^2 + \lambda w^T w \\ & = w^T X^T X w - y^T X w - w^T X^T y - y^T y + \lambda w^T w(3) \end{aligned}$$

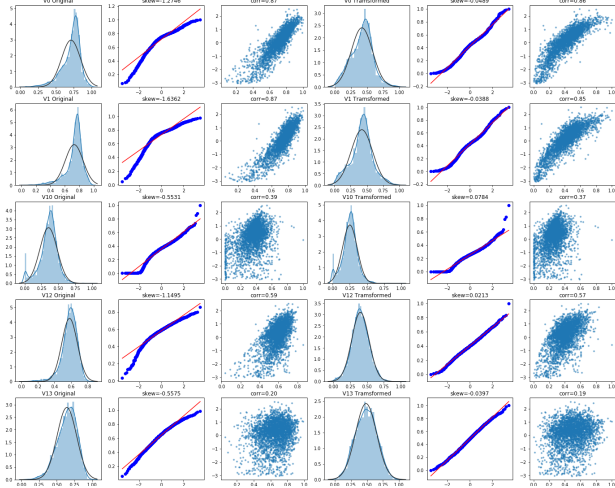


Fig. 3. Influence of Box-Cox transformation on the data distribution

$$\frac{\partial L(w)}{\partial w} = 2X^T X w - X^T y - X^T y + 2\lambda w = 0 \quad (4)$$

$$w = (X^T X + \lambda I)^{-1} X^T y \quad (5)$$

We apply Ridge Regression to detect outliers and remove them. The result is shown as Fig. 4.

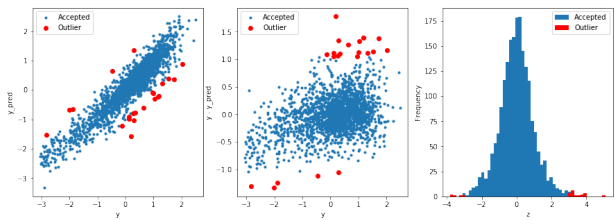


Fig. 4. Outliers visualization

One-class SVM is One-Class Support Vector Machine. Typically, the SVM algorithm is given a set of training examples labeled as belonging to one of two classes. One-Class SVM is an unsupervised algorithm that train on data that has only one class, which is the “normal” class. It infers the properties of normal cases and from these properties can predict which examples are unlike the normal examples. In this case, this module is particularly useful in scenarios where we have a lot of “normal” data and not many cases of the anomalies we are trying to detect.

Since we have eliminated most outliers by Ridge Regression, once again, we apply One-Class SVM to eliminate the rest few outliers.

3.6. Logarithm transformation

We do the logarithm transformation on target data to improve its normality. As Fig. 5 shows, the data is more normal.

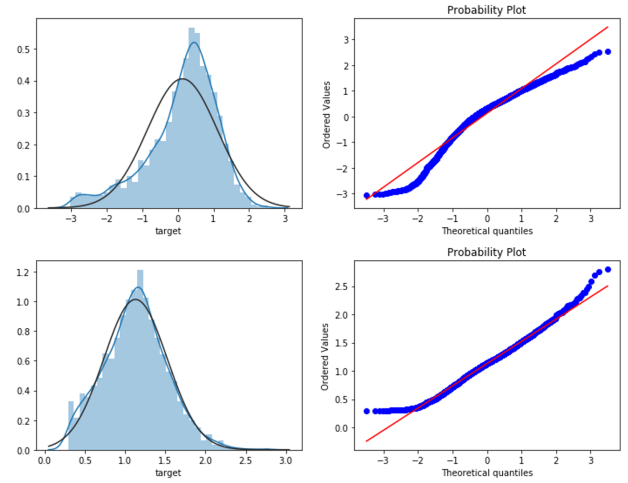


Fig. 5. Influence of logarithm transformation on the target data

After all the preprocessing above, finally we can achieve a clean training set. The training set now has dimension of (2621, 30), which indicates that only 2621 samples with 30 feature channels will be used in our future implementation.

4. METHODS

After the data preprocessing, now that the data is clean. We split the data into training and validation set.

Since the data are mostly under obvious regression tendency, we choose Regression models as the main way to predict our result. We evaluate the models based on their scores and mean squared errors. After couple experiments, here are the best four models: Linear Regression, Gradient Boosting Regression, Ridge Regression and MLP Regression.

4.1. Linear Regression

Linear Regression is one of the most widely applied models for prediction. It expects to establish an optimum line model to express the relationship between the features and outcomes.

$$y = X\beta + \varepsilon \quad (6)$$

4.2. Gradient Boosting Regression

Gradient Boosting Regression is a machine learning technique for regression problems, which produces a prediction model in the form of an ensemble of weak prediction models,

typically decision trees. It may increase its performance by decreasing the loss along the largest gradient.

We used huber loss, a loss function used in robust regression, which is less sensitive to outliers in data than the squared error loss. We use 100 estimators and set the learning rate as 0.06.

$$\hat{F} = \arg \min_F \mathbb{E}_{x,y}[L(y, F(x))] \quad (7)$$

$$\hat{F}(x) = \sum_{i=1}^n \gamma_i h_i(x) + const. \quad (8)$$

$$F_0(x) = \arg \min_{\gamma} \sum_{i=1}^n L(y_i, \gamma) \quad (9)$$

$$F_m(x) = F_{m-1}(x) + \arg \min_{h_m \in H} \left[\sum_{i=1}^n L(y_i, F_{m-1}(x_i) + h_m(x_i)) \right] \quad (10)$$

4.3. Ridge Regression

We have explained Ridge Regression above for eliminating outliers in data. We can still use it as model to train our data.

We set regularization strength alpha as 1, and fit the intercept for this model in calculations.

4.4. Multilayer Perceptron(MLP) Regression

A Multilayer Perceptron is a class of feedforward artificial neural network. MLP Regressor is a kind of supervised learning, which is known for its non-linear hiding layers between input and output. It is also reasonable for this project since the data is not under an overwhelming size.

We set the hidden layer sizes as 200, maximum iteration as 150. Early stopping is applied to avoid overfitting during the model training.

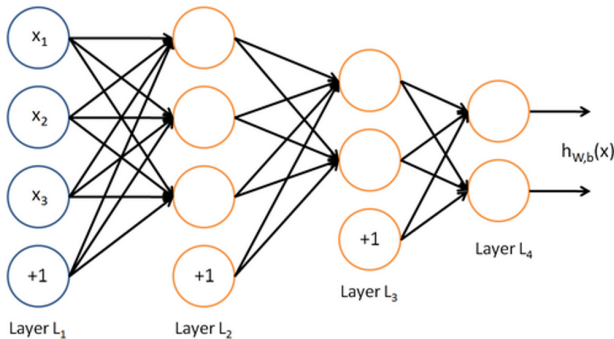


Fig. 6. An MLP Example

5. RESULTS

After all the model training, we predict on our validation set and achieve results as Table. 1. The result applies the score function to illustrate the accuracy and also implies the mean square error. According to these results, we can see that all these results achieve the outcomes with approximately the same score, while the MLP Regression has a relatively very small mean square error compared with the rest of the three models. But the MLP result might have a small variance in different training process, so we cannot say MLP is the best model in some way.

Linear Regression	
Score	0.8926894125981535
MSE	0.11053158993152493

Ridge Regression	
Score	0.8917557871111966
MSE	0.11149323884215377

Gradient Boosting Regression	
Score	0.8911242195088429
MSE	0.11214376338896286

MLP Regression	
Score	0.8930182569721768
MSE	0.11019287506302151

Table 1. Results

6. CONCLUSION AND FUTURE WORK

In conclusion, the models we use had almost the same score and mean square error on the validation sets. The Multilayer Perceptron Regression had a better performance among the models. To find which method is the best in this project, we need a larger dataset to validate those models and then draw a conclusion.

About future work, we will have a further investigation on the models to find optimal parameters that may contribute to a better performance. Besides, we will explore some other models that might also work. Last but not least, we will try different possible data preprocessing methods that might also contribute to our result, because we discover that among all the factors that may influence our result, data preprocessing plays the most significant role.

7. REFERENCES

- [1] Frederick M Steingress. *Low pressure boilers*. American Technical Publishers, 1986.
- [2] Yukun Ding and Yiyu Shi. Real-time boiler control optimization with machine learning. *arXiv preprint arXiv:1903.04958*, 2019.
- [3] MS Murshitha Shajahan, D Najumnissa Jamal, V Aparna, and MKA Ahamed Khan. Control of electric power generation of thermal power plant in tamilnadu. *Case studies in thermal engineering*, 12:728–735, 2018.
- [4] C Maffezzoni. Boiler-turbine dynamics in power-plant control. *Control Engineering Practice*, 5(3):301–312, 1997.
- [5] FP De Mello. Boiler models for system dynamic performance studies. *IEEE Transactions on Power systems*, 6(1):66–74, 1991.
- [6] Moustafa Elshafei, Mohamed A Habib, and Mansour Al-Dajani. Prediction of boilers emission using polynomial networks. In *2006 Canadian Conference on Electrical and Computer Engineering*, pages 823–827. IEEE, 2006.

8. CONTRIBUTIONS

Bolin He tried to visualize original data and implement data preprocessing. Besides, he tried to explore different possible models for implementation. Furthermore, he contributed to the presentation video editing and report writing.

Xiaoyuan Jiang gathered the materials as well as other necessary information for the group to understand the topic and complete the project. He also designed the outline of the proposal report and this one. Furthermore, he, together with the other group members, have completed some work on data processing.

Yidong Li made contribution to doing data preprocessing and using models to do prediction on the datasets and he also completed the final report, made presentation slides, recorded video of the presentation of the project.

9. REPLY TO REVIEW

Critical review from team 29:

How would nonlinear regression techniques perform without the preprocessing of the data?

Without data preprocessing, the score is below 0.7 which is much lower than the final result. Moreover, according to the principle of data processing, this loss of accuracy would be significantly amplified when analyzing large scale real-time data instead of validation. This is why we emphasize data preprocessing in length.

What prompted the original assumption that the input data is linearly related to the efficiency?

We did not assume the input data is linearly related to the efficiency. The interrelationship of each features were based on the investigation as well as discussion in the dataset section. That is why we try different models, and finally discover they are linearly dependent in some way, because the result of linear regression is close to other models.

Critical review from team 35:

Feature Extractions methods are not mentioned, I only saw some feature selections.

Actually, this project do not really include feature extraction which is optional according to the guideline of progress report. Refer to our dataset section. The data are some masking data collected by the boiler sensors with the collection frequency on the minute level. So we need not do feature extraction, because they are raw data collected by sensors.

Results section can show more comprehensive results rather than just MSE and score.

Yes, in some review researches, more complicated outcomes were extracted to determine the performance of the

models. In this project, however, we try to estimate the efficiency by training models. The best and only criterion is score and MSE, while score is also highly depends on MSE.

Critical review from team 69:

The data preprocessing is so extensive that useful features for the prediction may be disregarded. This could be potentially be solved by using other ML techniques for regression that would only require data normalization and anomaly exclusion.

This is a great idea, because ML is really good at this. However, we believe that useful features are not disregarded mostly by two factors. First, we set certain criterions to select features, while all the criterions make sense. Second, we can achieve less MSE and higher score by doing so, which means we are in the right direction.

Only using linear regression techniques most probably created an upper threshold to the accuracy that could be achieved.

Without datapreprocessing, the data and results are poor linearly dependent, that is why we emphasize data preprocessing at length. After our tremendous efforts on data preprocessing, our final data become much more linearly dependent. Finally, only with the results from different models, can we draw a conclusion that linear regression is optimal.

The data presented are from certain positions within the boiler or they change over the repetitions. If they are fixed why not use this information in the model selection to improve accuracy.

To be honest, the dataset is really a closed type and somewhat a black box. In other words, the real and physical name of the features are not accessible and can only be investigated in sequence numbers. Thus, although the data are collected from certain positions, we do not have the position information.