SUMMER INTERNSHIP REPORT

SUBMITTED BY

SWARNA LATHA S V

CHEMICAL ENGINEERING

102121072

NATIONAL INSTITUTE of TECHNOLOGY TRICHY

GUIDED BY

Dr. HIMANSHU GOYAL (Assistant Proffessor)

Department of CHEMICAL ENGINEERING

INDIAN INSTITUTE Of TECHNOLOGY MADRAS

CONTENTS

OBJECTIVE

To build a machine learning model for SERB project: “Data-assisted strategies to integrate detailed chemical mechanisms with reacting multi-phase flow simulations “.

ABSTRACT

Random Forest Regression Model can be built using Sklearn.ensemble Library which is used to train a data of size fifteen lakh. The data was taken at the temperature of 1073 K. The data was obtained from the Computational Fluid Dynamics (CFD) simulation. In Random Forest regression, the relationship between input features and the target variable is typically determined through the concept of "feature importance." Feature importance in a Random Forest is calculated based on how much the mean squared error (MSE) of the model increases when the values of a particular feature are randomly permuted while keeping other features unchanged. Features that have a larger impact on reducing the model's error will be assigned higher importance scores. The model works quite well with an R squared value of 0.9, which is a good performance.

INTRODUCTION

Random Forest makes use of a significant number of Decision Tree Regression in which every input feature or some of the input features to build a tree using the best split value by going through every unique feature of the input features, so that the mean squared error is minimized at that point. The algorithm does this in a repetitive fashion and forms a tree-like structure. The average value of every leaf node acts as a predicted value. The main parameters of the Random Forest algorithm are n_estimators, Max_depth, Max_features, Max_leaf_nodes , Min_sample_leaf, Min_samples_split, and so on. By optimizing the above-mentioned parameters, the model can obtain a better accuracy. The accuracy of a model is measured either in terms of loss or R squared value. I have used R squared metric to measure the accuracy.

KEYWORDS:

- n_estimators: No. of Decision Tree regressor
- max_depth: The depth of each Decision Tree
- max_features: Maximum number of features considered for splitting at each node
- min_leaf_nodes: Minimum no. of leaf nodes present at each Decision Tree
- min_samples_split: Minimum number of samples required to split an internal node in the decision tree.
- min_samples_leaf: Minimum number of samples required to be at a leaf node.

FEATURE SELECTION AND DATA CLEANING

a. Feature Selection:

As chemical reaction rate of a specie depends upon the concentration of species participating in the chemical reaction, it is mandatory to select every species as input features.

On Visualizing the dataset, it is noticeable that few species' mole fraction against their respective rate is zero at all the instances, found five species namely,

- CH2OH
- C2H5
- CH2CO
- CH2CHO
- AlCH2-C7H

are idle with zero mole fraction and zero reaction rates, so removed those columns with their respective reaction rate columns which are called as quasi steady state species.

b. Data Pre-processing:

1) The datatypes of all the attributes were checked, found two columns with object datatype and converted to floating numbers.
2) Reduction of the dataset: Removed some of the rows containing mole fraction of nitrogen equal to 1.0, as the sum of mole fraction of species participating in a chemical reaction should be equal to 1 and it is quite visible that species other than Nitrogen are not participating at that time instant, so we can conclude that these species don't play any role in the chemical reaction at the particular instance, by doing so the dataset got reduced to 3.4 lakh.
3) Removed the datasets with negative mole fraction due to the same reason mole fraction value cannot be in negative value, so the size of the dataset got reduced to (340237, 80).
4) Divided the datasets into two parts: training dataset with size of 300000 data, testing dataset with size of 40237 data, shuffled the data though Random Forest can handle the data without shuffling.
5) Min- Max scaling was applied to every input feature as columns of the dataset are non-uniform and vary exponentially from each other.

CODE IMPLEMENTATION AND MODEL OUTPUT

Built a Random Forest Regressor from the sklearn.ensemble library and to evaluate the model's performance, sklearn metrics library was used, model fitted quite well for the given dataset with a short time span of 2 minutes 48 seconds. The accuracy of model which was obtained as follows.

- Mean Squared Error: 1.2795695371e-05
- R-squared : 0.912069281

However, some of the rates of individual species' R-squared values are highly inconsistent and deviating. Some of them are in negative R squared value.

The species are:

	SPECIES	R-Squared Value
1.	HCO:	-4.409529
2.	CH3:	-1.415212
3.	C2H3:	-2.663814
4.	CH3O:	-2.354861
5.	OH:	0.221891
6.	H:	0.254602

The R squared values for OH and H are in the range of 0.2 . The species other than the species mentioned above fit well into the Random Forest Regressor Model with R squared value around 0.998 and Nitrogen is an exception from this list as its R squared value is 0.935. The efficiency of the model is disturbed by these six above-mentioned species. Random Forest has given a desired output without even tuning its hyperparameters. However, I used GridSearchCV method of hyperparameter tuning and got the same result.

I retrieved the parameters involved in the Random Forest Regressor model which was performed to obtain the result and saved it in a python dictionary. I plotted a parity plot for every species' chemical reaction rates.

ADVANTAGES OF RANDOM FOREST REGRESSOR

The most significant advantage is Random Forest Regressor is not as time consuming as the Artificial Neural Network (ANN). Random Forest Regressor merely took two minutes to fit the whole dataset and for predicting the targets of test dataset.

Random Forest does not have as many hyper parameters as ANN does, so it is easy to tune the hyperparameters of Random Forest Regressor. In most cases Random Forest does not need any hyper parameter tuning as it works with splitting with the help of the best values out of every value.

It is feasible to code rather than coding the Artificial Neural Network which requires intense coding.

DRAWBACK OF RANDOM FOREST REGRESSOR

The weights of every species' mole fraction participating in the chemical reaction and the bias value cannot be obtained as it is retrieved from ANN model.

CONCLUSION

The Random Forest Regressor shows good accuracy in predicting the output, and it consumes very little time to perform the whole process of training and predicting in minutes. Though the algorithm uses less computational time, it has its own disadvantage.

The ultimate goal of building the machine learning model is to develop an equation relating the correlations between the species participating in the reactor and its chemical reaction rates, which

cannot be computed easily. Usually, the chemical reaction rate of a species of any order higher than two cannot be computed manually and even sophisticated systems like CFD takes so much time and it is computationally expensive.

The R squared values of six species are very low which might be due to the very diminished values of some columns which are exponentially low compared to other species even after scaling. The data of some of the columns are scattered drastically, such as the normalized LVG-C6H10O5, where a quarter of the data fall below 1.03e-56, while on the other hand three fourth of the data fall below 2.52e-14. Though the values are significantly small, there is a huge deviation between the two values. As the rates of these species may depend on those columns mainly according to their respective reaction rates. For better performance, we can start working on these elements' data by applying special scaling techniques or careful removal of such outliers would be better choice to obtain a good result using any of the machine learning model.

I have attached the file for parity plot, please do consider having a look at it.

Parity plot of rf.png