Supervised Learning Approach on Optimizing Hydrocracking Yield for Profitable and Sustainable Production

Ziyang Lin

Abstract

This reports outline the approaches taken to select a best-performing supervised learning models that study the hydrocracking process to determine the most influential control variables of the reactor on the yield of effluent hydrocarbon mixture in the target range of densities, and to predict the yield given control variables. Higher target range yields are desired to attract a higher profit on sale and also reduce production waste. The results of these models are interpreted in terms of their identified influential control variables, and quantify their effects in improving the target range yield. In particular, we compared linear regression, lasso regression, random forest regression, and regression tree boosting on the root mean squared error on the test set and discovered lasso regression to be the best performing model. The models all have relatively low records of RMSE, and we saw that catalyst type, reactor temperature, reaction time, and feed hydrocarbon mixtures on the same and nearby intervals as the target range all have a positive impact on target range yield.

1 Introduction

Hydrocracking is an industrial process to break down hydrocarbons into smaller molecules, which is used to increase the proportion of extracted hydrocarbons of shorter molecules. These molecules attract a higher profit on sale as they are typically more useful to consumers. The most useful and valuable hydrocarbons are those within a density in a specific, intermediate range known as the target range.

The effluent hydrocarbon mixture composition is controlled by adjusting the temperature of the reactor, the catalyst used, and the time for which the hydrocarbon mixture is within the reactor. It also depends on the composition of the feed mixture. In this report, we will take a data-driven approach to understand control variables that are most influential on the effluent hydrocarbon mixture yield of the target range and use them to make predictions. This initiative will not only increase profitability, but also make the process more sustainable as it minimizes reactor output waste.

We will apply different supervised learning methods to build models that explain the target range yield and the control variables. In particular, we will look into both parametric and non-parametric approaches and examine model performance of each on the test set using the same metric. We will start from exploratory analysis that aims to investigate the distribution of the target range yield, and its correlation with the control variables. We then use these preliminary findings to choose suitable supervised learning models. We aim to determine the most influential factors among these variables, and also select a best-performing model to make predictions and use the fitted model to provide advices for the production team.

1.1 Exploratory Analysis

The data set used in this study contains 497 observations of 47 variables. The observations are daily records from 2020-10-15 to 2022-02-23 for both the control variables, and the composition of effluent hydrocarbon mixture yield. There are two types of control variables, 3 of them represent the reactor settings which includes the type of catalyst used as a categorical varible of 3 levels, the reactor temperature in degrees Fahrenheit, and the reactor residence time of the mixture in hours. The other type is the composition of the feed hydrocarbon mixtures, where the proportion of the overall mass in each of 20 density intervals is given as 20 individual numerical variables. The first interval represents the longest, heaviest molecules and the last represents the shortest, lightest molecules. The remaining 20 variables are the proportion of the overall mass in the 20 density intervals for effluent hydrocarbon mixture composition. We are only interested in the target range yield, which is the combined yield of interval 13, 14, and 15 for effluent composition.

The data set is complete and clean at the first glimpse without any missing observations, so no pre-processing is needed except for calculating the combined yield for the three target intervals. First we consider the empirical distribution of the combined target range yield (Figure 1), we see that the yield roughly follows a Gaussian distribution without obvious skewness and outlying observations. This suggests that methods assuming a Gaussian response such as linear regression may be suitable.

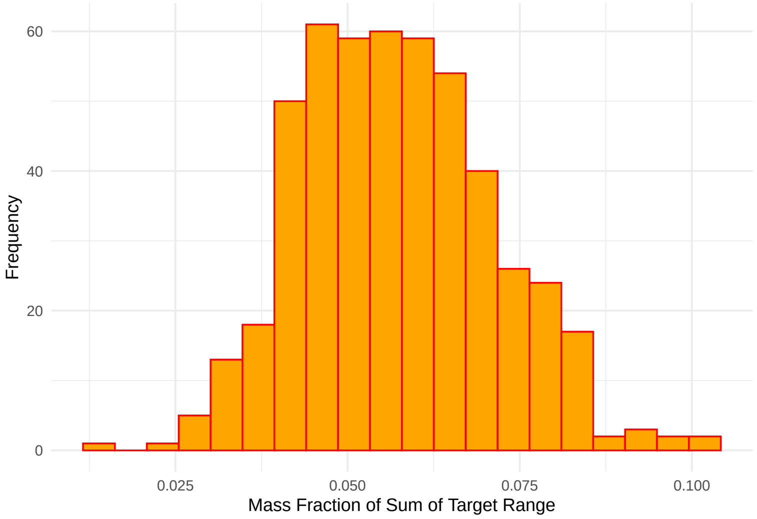

Figure 1: Empirical Distribution of Sum of Target Range Densities

In fact, when we consider the relationship between reactor setting variables and combined yield, we see there is a distinctive linear pattern with positive correlation (Figure 2). There is also a visually distinctive difference in means for different catalyst types (Figure 4, Appendix). This further justifies the adequacies of linear models. We also have reasons to believe that the feed composition in the corresponding intervals (13, 14, and 15) would have correlation with the combined yield of the target range (Figure 2). These findings would give us a good starting point for model choosing and variable selections.

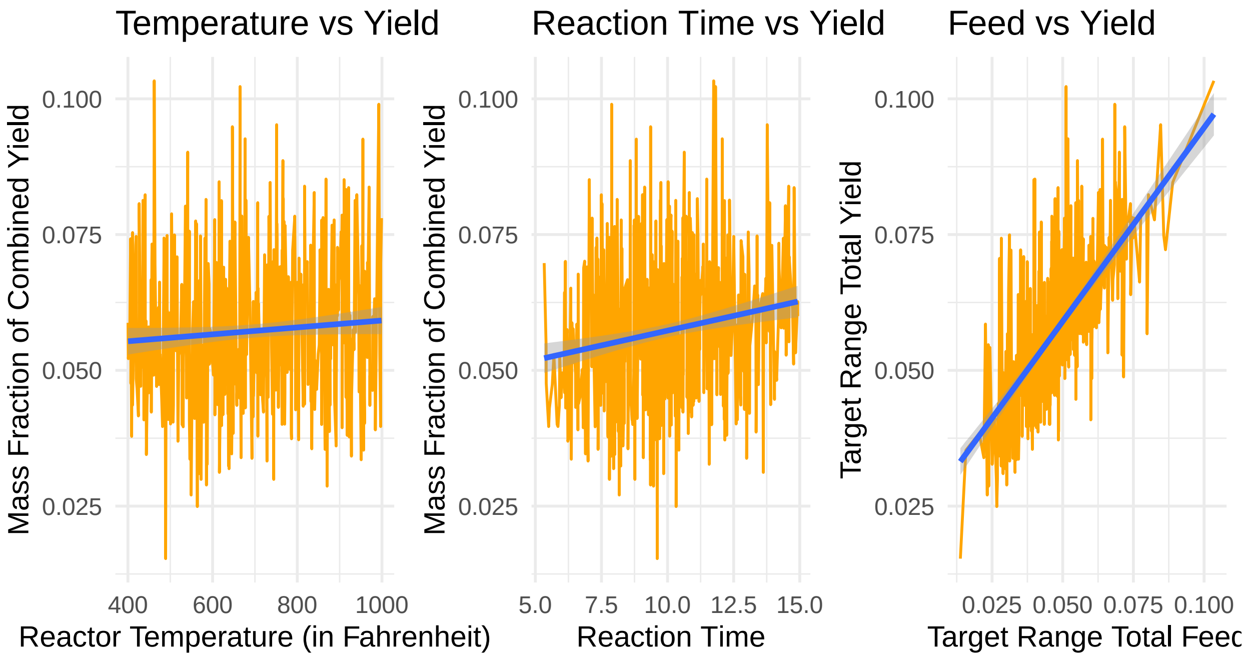

Figure 2: Some Control Variables versus Combined Yield

2 Methods

2.1 Parametric Approach

Before fitting models, we should first divide the data into training and test set for model fitting and out-of-sample prediction errors to compare models. We will take 70% of the observations as training data and the remaining as testing data. Since our response is a continuous variable, we should use regression methods for model fitting.

As the exploratory findings suggest, a linear model or its variant may be suitable for this data set. We will first consider the simple linear regression as a parametric approach to fit a model on the control variables to model the conditional mean of the combined yield. The model will take the form $Y_i|\boldsymbol{X_i} = \boldsymbol{X_i}\boldsymbol{\beta} + \epsilon_i$ where $\epsilon_i \sim N(0, \sigma^2)$ such that σ^2 is constant across all observations. This model is fitted by minimizing the loss function of residual sum of squares, and it will produce $\boldsymbol{\beta}$ coefficient estimates that allow us to quantify the effect of each control varible and whether it differs from 0.

However, fitting a full linear model may also include unimportant predictor, we will therefore apply two methods to potentially obtain a model that contains most important variables and also generalize to new data well. We will use stepwise selection using the AIC criterion, which is a likelihood-based statistic to trade-off model complexity and goodness-of-fit, to get a model with only important predictors. We will also introduce a Lasso panelty term to the full model, so that the modified loss function is $\ell_{\text{lasso}}(\beta) = \sum_{i=1}^{n}(y_i - X_i\beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$ where the parameter λ that specifies the strength of penalty will be tuned using cross-validation.

2.2 Non-Parametric Approach

We will also explore how non-parametric approaches perform for our problem by considering tree-based methods. In particular, we will to fit a random forest regression model and a boosting regression tree model to explain the variability of the combined yield by the control variables and generate out-of-sample predictions.

Unlike linear regression models, trees do not take the approach of estimating a set of parameters, but instead it is constructed by successive binary splits in predictor space. A regression tree estimate the model function by $\hat{f}(x) = \sum_{m=1}^{M} \hat{c}_m I(x \in R_m)$ where $R_1, \cdots, R_M$ are regions in predictor space, and c_m is a constant for the predicted response of all observations in any region R_m. The regression tree is also optimized using the residual sums of squares, but unlike linear regression, it predicts a new data not by the estimated coefficients, but by the predictor space region that the new observation falls into.

Trees are high-variance models in general, but we can use bagging to draw some number B of bootstrap samples from the training data, and fit a tree to each of them, then aggregate the prediction by taking the mean of predictions from all trees. Random forest is a variation to this method where we use a random sample of $m = \sqrt{p}$ predictors to determine each split where p is the total number of predictors. This way the variance of the model can be largely reduced, and random forest can also be interpretable when considering the reduction in MSE when a split uses a certain variable. Boosting is also a technique to improve performance of tree models by building trees sequentially such that one depends on the other instead of independent samples as used with bagging.

We will first fit a random forest regression model for combined yield using all available control variables, adn evaluate its performance and interpet the predictors' effects. Finally, we apply the boosting algorithm to see if this gives an even better result.

Since we are comparing models of different constructions, it is important to set up a uniform metric for comparison. As these are all regression methods, we will compare them by the square-rooted mean squared error $RMSE = \sqrt{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2/n}$ on the out-of-sample test set which contains 30% of all observations. We will also compare the selected useful predictors by each method and use these results to drive the advices to the production team.

3 Results

We start by fitting a linear regression model with combined yield being the Gaussian response, and all control variables except feed fraction at interval 20 as the predictors. We get rid of one interval fraction because we know the mass fractions of 20 intervals sum to 1 so we only have a degree of freedom of 19. The AIC-selected model only excludes a few feed fraction variables but according to the t-statistic of each coefficient estimate, it identifies some control variables to be more important (Table 1).

We see that the feed fraction at interval 12, 13, 14, and 15 are very influential on the target range yield. We also see that both reactor temperature and reaction time have positive effect on the yield, and that catalyst type 1 and 2 both have higher expected yield than catalyst 0 when holding other factors fixed. In general, this model has a good fit with $R^2 = 0.8211$, and the $RMSE$ on the test set is reported at 0.0062.

Table 1: Predictors with Strongest Effects

	Estimate	Std Error	t-Stat	Lower 95% CI	Upper 95% CI
feed_fraction_13	0.8888	0.0401	22.1699	0.8100	0.9677
feed_fraction_14	0.7654	0.0430	17.7951	0.6808	0.8500
feed_fraction_12	0.5730	0.0335	17.1125	0.5072	0.6389
catalyst2	0.0099	0.0008	12.3633	0.0084	0.0115
feed_fraction_15	0.4443	0.0446	9.9715	0.3567	0.5320
catalyst1	0.0066	0.0008	7.9705	0.0050	0.0082
through_time	0.0011	0.0001	7.1479	0.0008	0.0014
feed_fraction_11	0.1967	0.0301	6.5356	0.1375	0.2560

Now, we can take the variable selection step further by including a penalty term to perform a Lasso regression where the strength of panelty λ is tuned by cross-validation within the training set. With a grid of 100 λ candidates evenly splitted from $\exp(-11)$ to $\exp(-4)$ based on the shrinkage plot (Figure 4), the cross-validation outputs $\lambda_{\text{min}} \approx 0.000113$ as the one that produces the lowest $RMSE$ on the test set at 0.0061, which is slightly lower than linear regression. The Lasso regression gets rid of feed fraction at interval 4, 6, 17, and 18, which is comparable with the AIC-selected linear regression model. In general, we see that parametric regression models with Gaussian response perform well for our data set in terms of out-of-sample errors.

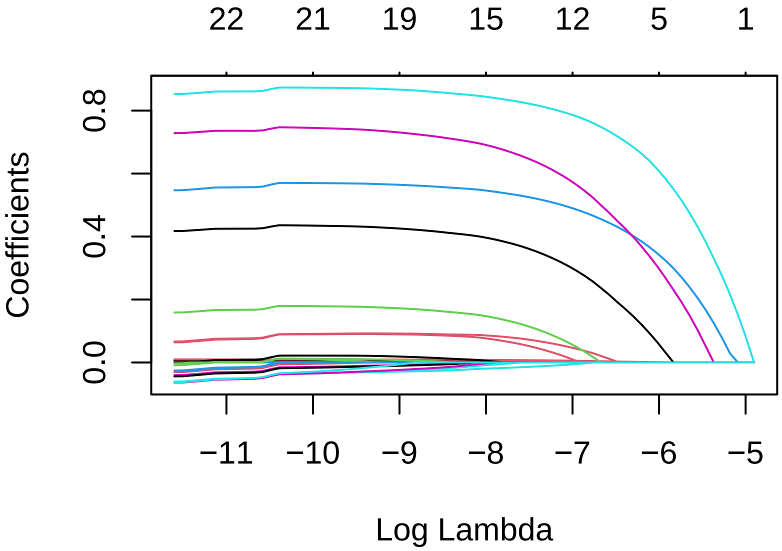

Figure 3: Shrinkage Plot of Lasso Regression

We next fit a random forest regression model as a non-parametric approach to our problem. With this method, each tree in the bootstrap aggregation only select a subset of 7 predictors from all 23 predictors. The test set $RMSE$ is 0.0093 for random forest which shows a drop from the parametric regression model performance. We can also see that the random forest captures similar feed fraction intervals, catalyst, and reaction time as important predictor, but the effect of temperature is not as strong as the parametric models.

We finally implement a regression tree boosting model with cross-validation to select the optimal number of trees in the model. Boosting algorithm have three hyperparameters to tune, and they are

Table 2: Boosting Tuning Parameters and Results

lambda	d	B	RMSE
0.01	3	1000	0.0069021
0.01	5	997	0.0068768
0.05	3	539	0.0068493
0.05	5	334	0.0069669
0.10	3	240	0.0068510
0.10	5	113	0.0072398

Table 3: RMSE for Model Comparison

Model	RMSE
Linear Regression	0.0060
Lasso Regression	0.0061
Random Forest	0.0093
Tree Boosting	0.0068

a shrinkage parameter λ, the maximum depth of any one tree d, and B the total number of trees. We manually specify $\lambda \in \{0.01, 0.05, 0.1\}$ and $d \in \{3, 5\}$, and then use cross-validation to tune B, which gives the below hyperparameters and test set $RMSE$ (Table 2). In general, these models perform comparably well, and the best model has $RMSE = 0.00685$ which is still lower than the Lasso regression

Now we can put all $RMSE$ from the four models together with the best-performing boosting tree model, and observe that Lasso regression has the lowest record (Table 3).

4 Summary

4.1 Conclusion

From the above model fitting results, we see that all four models are able to capture similar set of influential control variables and generalize well to unseen test set data. These would give us confidence in concluding the important factors that affect the target range yield of effluent hydrocarbon mixtures, and in predicting yields given these control variables for future data.

In particular, we found that parametric approaches performs better than non-parametric approaches in terms of out-of-sample errors, among which Lasso regression holds the top performance. From linear regression and random forest results, we see that the type of catalyst used, reactor temperature in degrees Fahrenheit, reaction time, and feed hydrocarbon mixtures on the same and nearby intervals as the target range are all influential variables to target range yield. In particular, they all exhibit positive correlation with the yield, which means increasing any of them will increase the expected mean combined target range yield of effluent hydrocarbon mixtures.

4.2 Limitations

Besides useful insights, there are still drawbacks in our approaches and some future steps to take to provide more reliable results. First, since the feed fractions in 20 intervals sum to 1, increases in some intervals will cause decreases in others, which means we may have multicollinearity in these variables. When linear regression models contain multicollinear variables, coefficient estiamtes could have the wrong sign and larger variance. This could potentially impact the model's predictive ability.

When we do cross-validation with boosting, we only compare six of the hyperparameter combinations, but their dependence could be more complex, and we might have avoided a better model. Additionally, the $RMSE$ metric mainly concerns the prediction error, but does not take variance of predictions into account. We could also integrate the variance of predictions as another metric so we could select a model with high accuracy but also low variability. Finally, except for the four models, we could have a wide range of other supervised learning methods to consider such as regression with splines, and principal component analysis.

5 Appendix

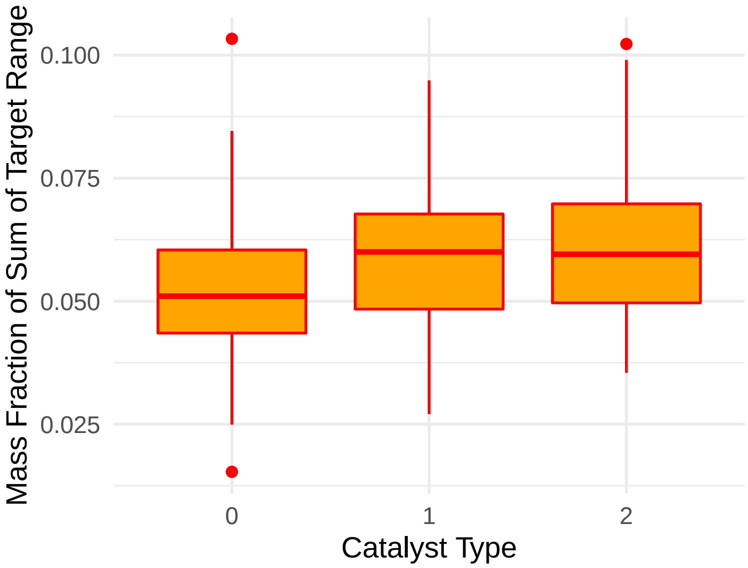

Figure 4: Catalyst Type versus Combined Yield

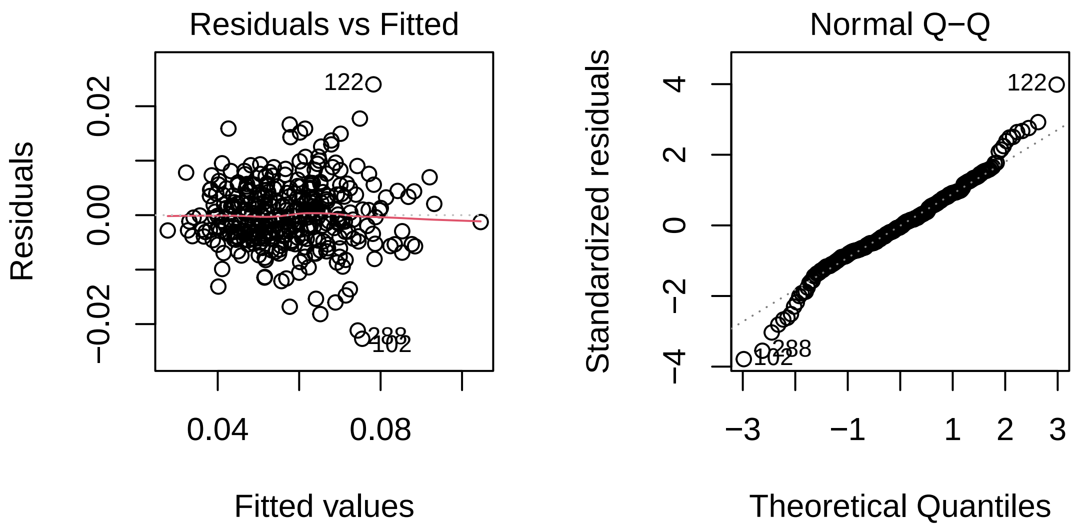

Figure 5: Diagnostic Plots of AIC-Selected Linear Regression Model

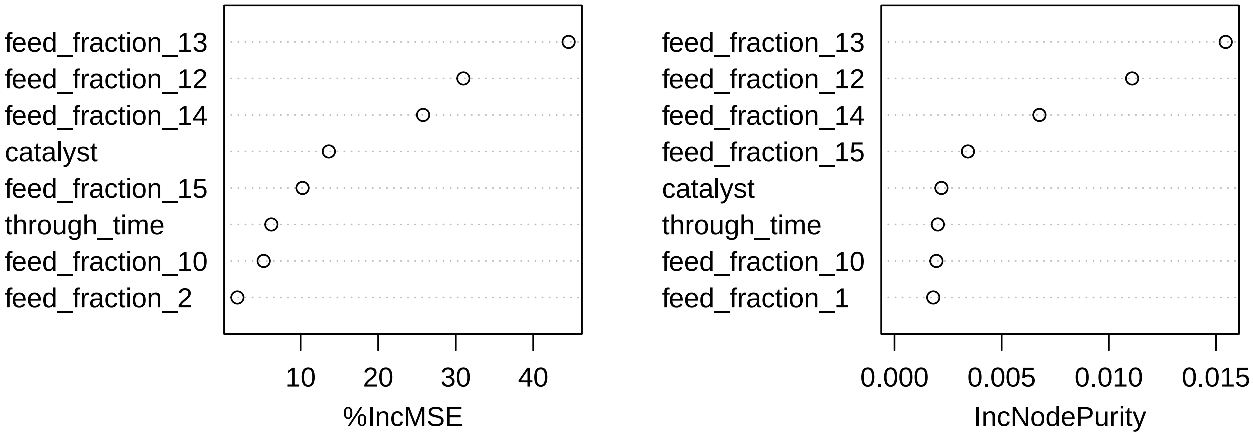

Figure 6: Variable Importance Plot for Random Forest