Because of the limitation of the length of the report, I think we only need to select several models to explain in the report. Maybe we only need one person to write this part of the report and select models, which will make the report more integrated.

In order to help us select models and write the final report, I will explain and introduce my two models here. First, I will make a brief introduction to my model and analyze the results of the model and my conclusions. Finally, some references I found are given.

## Linear Regression Model

### Model introduction

Linear regression attempts to model the relationship between a scalar response and one or more explanatory variables. In this project, the relationship between Response variable and Group\_x variables is the focus of our research. Multiple linear model is a good choice to explore the relationship between them. When building the model, we mainly focus on three aspects: R-square, p-value and VIF. R-square represents the degree of model fitting. P value represents whether the variable is statistically significant in the model. And VIF represents whether the variables in the model have collinearity.

Multicollinearity mainly affects the regression coefficient and statistical significance of a single independent variable, but usually does not affect the overall prediction performance of the regression model, including the prediction accuracy and goodness of fit statistics. Therefore, if the main purpose of constructing regression is to comprehensively consider the value of multiple independent variables to predict the value of response variables, and regard the fitted regression as a predictive model without explaining the contribution of a single independent variable, the multicollinearity problem can also be ignored.

But in this project, we need to find several variables that have the greatest impact on the product, so we must focus on collinearity. The method of manual selection and automatic selection is mainly used in the establishment of linear regression model. When removing high collinearity variables, we should pay attention to the changes of R-square and P value in the model, and it is ideal to balance their changes in the model.

### Results

Here I will summarize my findings and the results of my model.

Here, we try two kinds of data sets, one is the original data set, the other is the normalized data set. And the results of the model will be given below. As we can seen from Figure 1 and Figure 2, there are some differences in variables, but the method of selecting variables is the same, except that the data sets are different. I think it is largely because the distance between data becomes shorter after normalizing.

From the estimate column of linear model, we can find that the some coefficients of variables are positive and others are negative. The variable Group1-11 has a great positive impact on the Response. This means that as the value of Group1-11 increases, the value of response will rise quickly. In the composition of the product, we can try to increase the amount of Group1-11, which can improve the probability of the product passing the test.

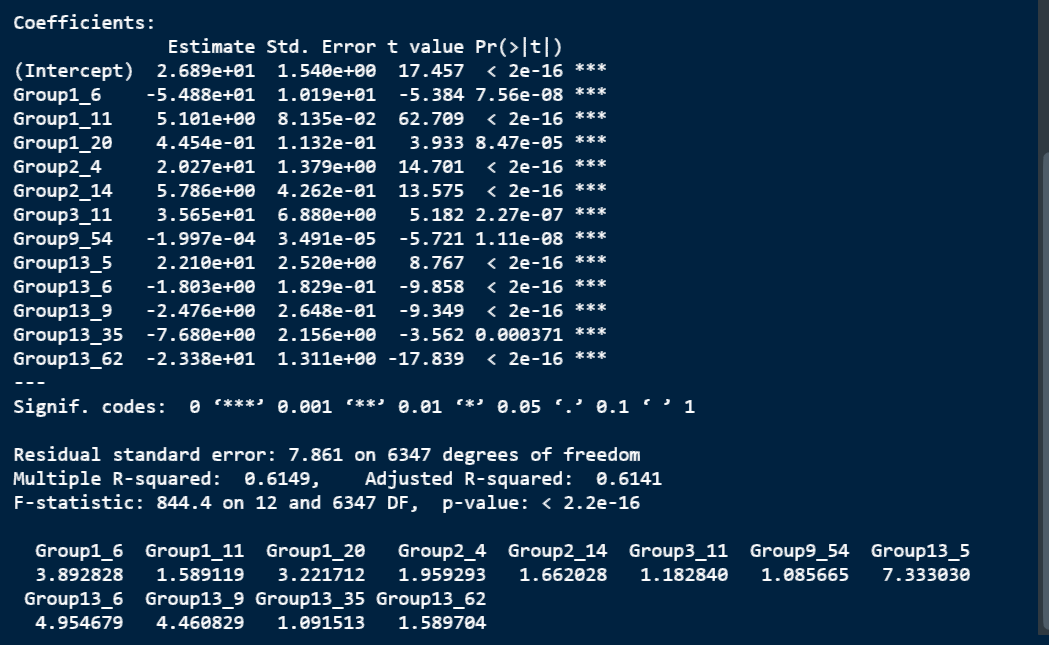


Figure 1 original data set

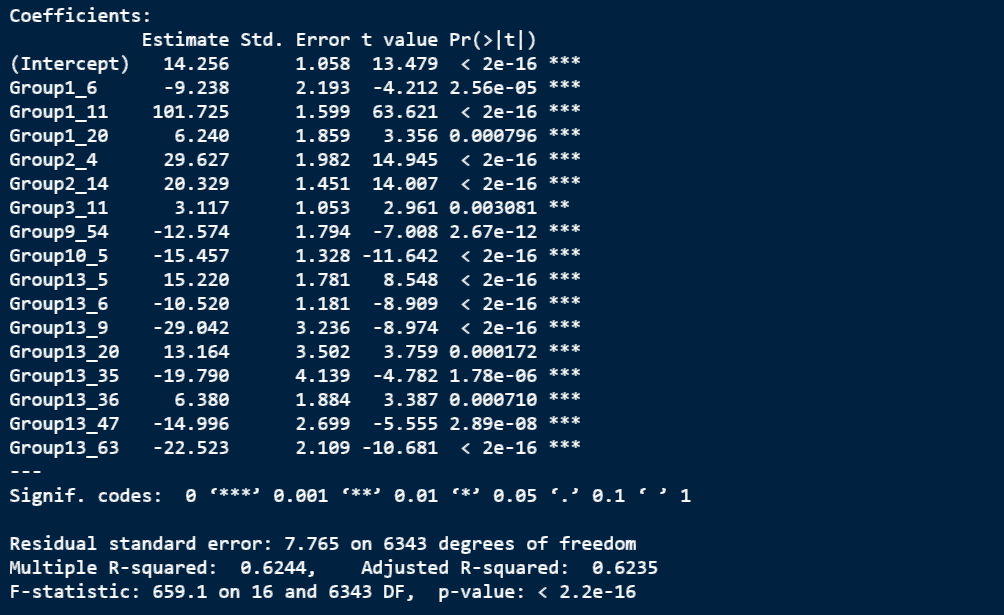


Figure 2 Normalized data set

At the same time, the linear model also obtains some variables that have great influence on Response. Group13 had the most subgroups with significant effect on Respone. And we can also list the linear relation of Response.

we compare the prediction ability of the model by looking at the proportion of the value of Response which more than 50. It can be seen from the Figure 3 that some values exceed 100, which proves that the data ser has some values that can not be well predicted. And only 8.7% of the products passed the test. Figure 4 shows the Response distribution of the original dataset. 18.7% of the products passed the test.

In addition, I verified some properties of the model in the code, such as homoscedasticity and linearity. The results of this part are in the code, which is not shown here.

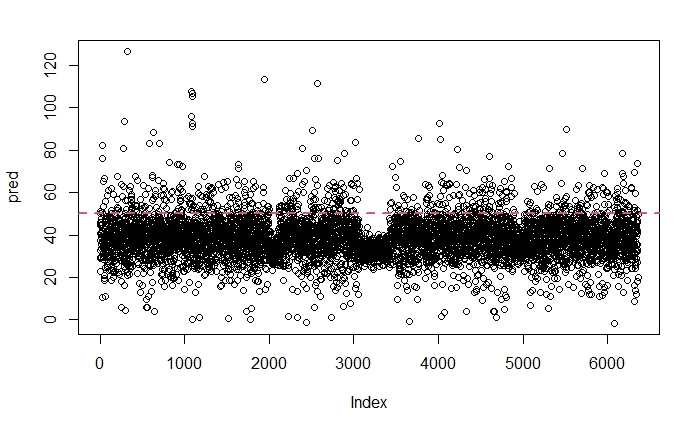


Figure 3 Response distribution of model prediction (Linear)

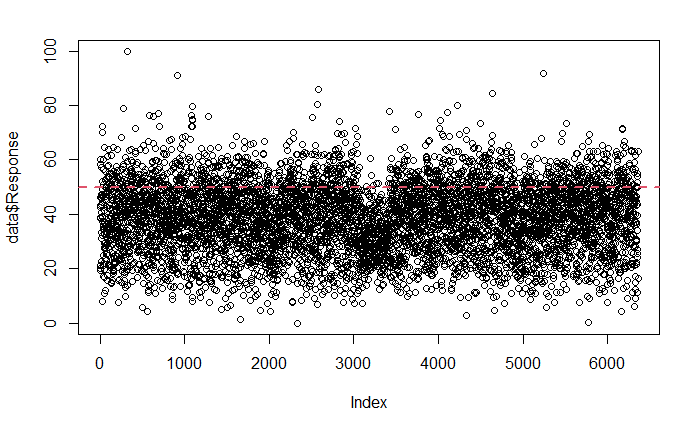


Figure 4 Response distribution of original data set

## Random Forest (Classification model)

### Model introduction

At the same time, the classification model is established to find the most important variables that affect whether the product can pass the test. Here, if Response exceeds 50, it is judged as pass (1), and the rest is fail (0). The random forest is used to establish the classification model.

Random forest is a supervised learning algorithm and an extension of decision tree. It is not sensitive to multicollinearity, and the results are robust to missing data and unbalanced data. The working process of random forest can be summarized as follows:

1. Assuming that there are N objects and M variables in the training set, N objects are randomly selected from the training set to construct a decision tree (The selected objects will be put back to the training set);
2. In each node, m < M variables are randomly selected as candidate variables to segment the node, and the number of variables at each node should be the same;
3. All decision trees are generated completely without pruning (the minimum node is 1);
4. The process of (1) - (3) is repeated to obtain a large number of decision trees; The class of the terminal node is determined by the mode category of the node (similar with DT).
5. For new observation points (test data set), all trees are used to classify them, and the classes are generated by the majority decision principle.

Compared **with** other classification methods, random forest usually has the following advantages:

1. The accuracy of classification is usually higher;
2. It can effectively deal with data sets with high dimensional features (multivariate), and does not need dimension reduction;
3. It also has advantages in dealing with large data sets;
4. **It can be applied to data with a large number of missing values**. (I think this advantage is very suitable for the data set of this project)
5. **It can measure the relative importance of variables to classification at the same time.** （It can help us to screen out the most important variables）

Disadvantages:

1. When the number of decision trees is large, the training time is long.
2. When the noise in the data is large, the model may have over-fit problem.

### Results

For this project, our random forest model has also achieved good results, and the prediction accuracy is **90%**. At the same time, we also find some variables that are most important to classification. It should be noted that the results of the random forest model will change with different seeds. In general, the performance of the model is kept in a certain range. The data set used here is the normalized data set. Overall, the model is not affected by the change of data set, only some variables have some differences in importance ranking

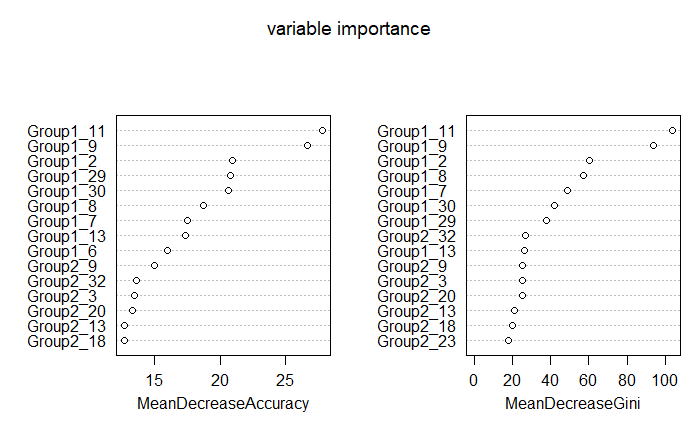


Figure 5 Classification model (normalization)

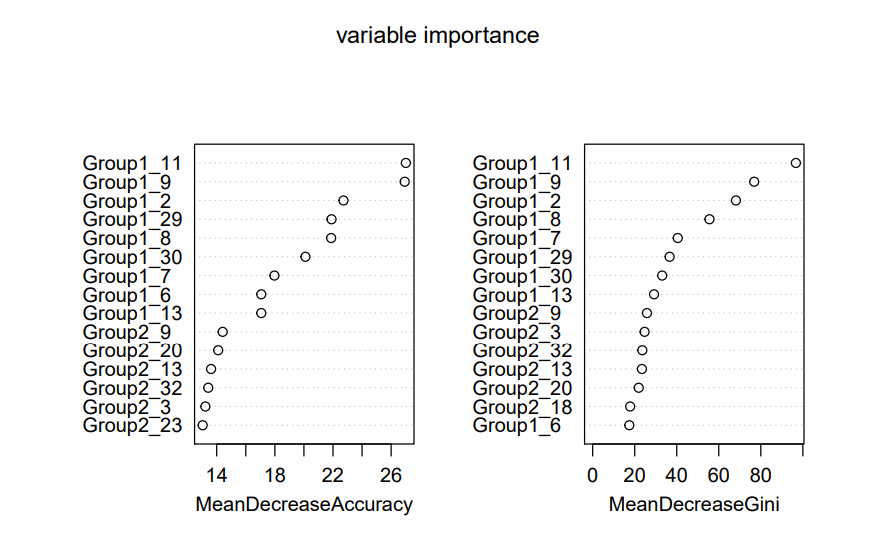


Figure 6 Classification model (origin)

**Mean Decrease Accuracy (%IncMSE) -** This shows how much our model accuracy decreases if we leave out that variable.

**Mean Decrease Gini (IncNodePurity) -** This is a measure of variable importance based on the Gini impurity index used for the calculating the splits in trees.

The higher the value of mean decrease accuracy or mean decrease gini score, the higher the importance of the variable to our model.

Group1\_11, Group1\_9 and Group1\_2 are the most important variables that affect whether the product passes the test. The top 15 variables of importance were concentrated in the Group1 and Group2. A classification model can reflect that we are exploring different ways of data analysis.

## Random Forest (Regression model)

Random forests can also be used to build regression model. Here, random forest fit the relationship between Group variables and Response.

### Results

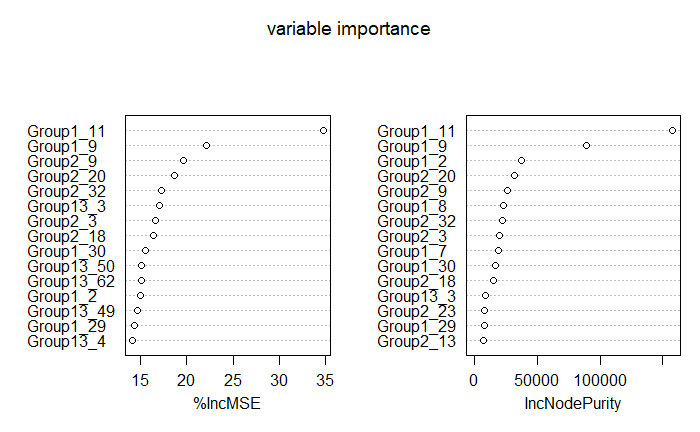


Figure 7 Regression model ((normalization)

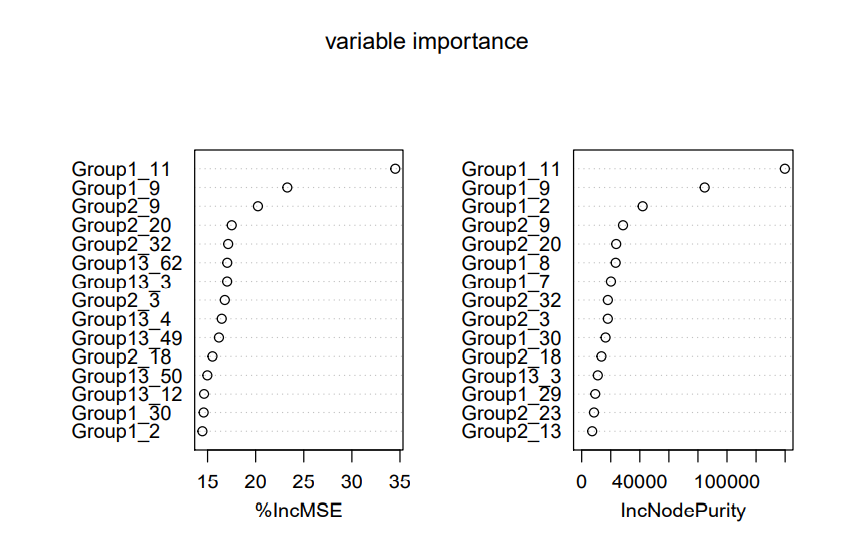


Figure 8 Regression model (origin)

When using %IncMSE as the standard, the top three variables of importance are Group1\_11, Group1\_9 and Group2\_9. When using IncNodePurity as the standard, the top three variables of importance are Group1\_11, Group1\_9 and Group1\_2.

To sum up, Group1\_11 and Group1\_9 are also the most important variables in regression model. The top 15 variables of importance were concentrated in the Group1, Group2 and Group13. Because random forest is not a linear dimension reduction method, its regression model results will be different from the linear regression results. But by comparison, we can find that Group1\_11 is of great importance to Reponse in both models. At the same time, the performance of the model is also very good (% Var explained: about 69).

Here we try to build a new regression model with only the top 15 variables of importance, and finally find that the performance of the new regression model is not as good as the original model. In addition, the performance of the model constructed by the first 15 variables based on %IncMSE is better than that based on IncNodePurity.

I think Figure 10 is a good random forest regression model. When the number of variables is greatly reduced, the %Var explained does not decrease much.



Figure 9 RF model composed of all variables

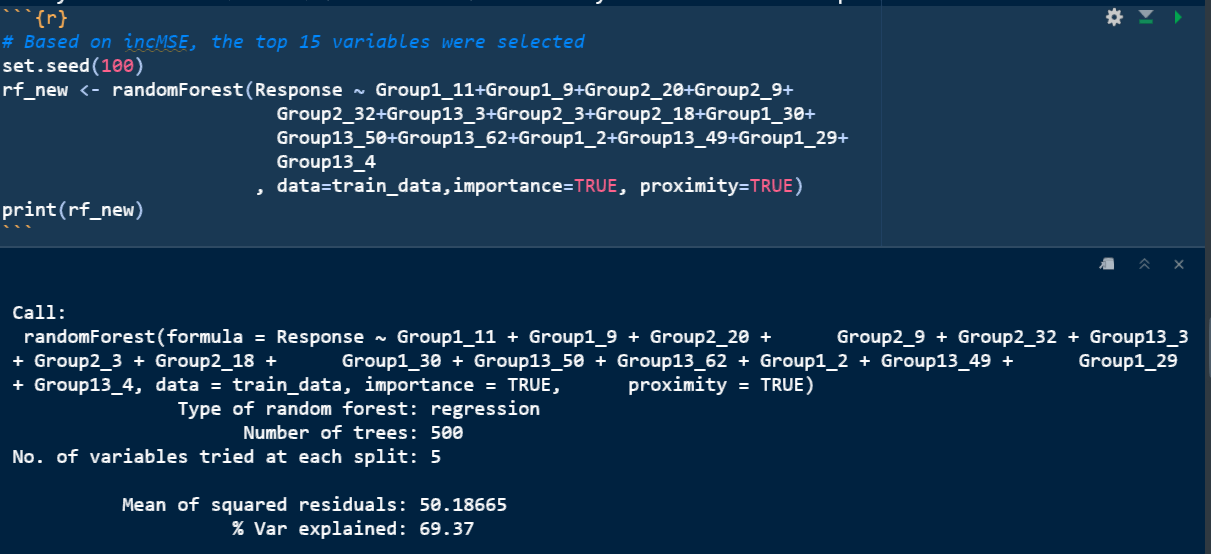


Figure 10 RF model composed of top 15 important variables (%IncMSE)

In the same way as the linear regression model, we input the model into the test set and observe the distribution of Response. 15.26% of the products passed the test. Compared with the linear model, it is closer to the data distribution of the original data. At the same time, we found that the predicted data had no outliers. Figure 10 shows the Response distribution of the original dataset. 18.7% of the products passed the test.

Compared with the linear model, the result of random forest model is more reasonable, and the distribution of response is closer to the original data set.

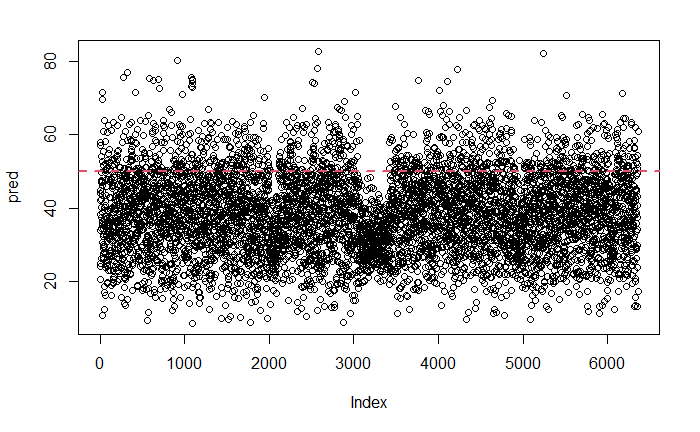


Figure 11 Response distribution of model prediction (RF)

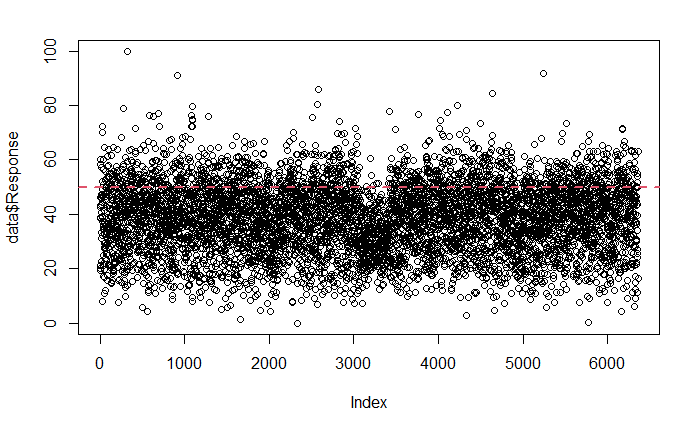
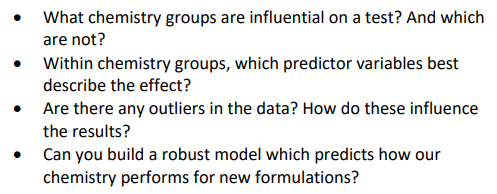


Figure 12 Response distribution of original data set

## Conclusion

Through the above models, we can find the Group and subgroup variables that have the greatest impact on Response. I personally think these two models fit our project requirements very well.



## Reference

<https://www.rdocumentation.org/packages/randomForest/versions/4.6-14/topics/randomForest>

<https://pubs.acs.org/doi/10.1021/ci034160g>

<https://www.sciencedirect.com/topics/engineering/random-forest>

<https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm#inter>

## Hint

I didn't find a good reference source for the linear model. The main reference is the content of the previous course. If there is anything unclear, please let me know and I will continue to add some content.

**My modified code has been uploaded to GitHub (latest version).**