基于集成学习的乙醇偶合制备 C4 烯烃问题研究

摘要

C4 烯烃广泛应用于现实生产生活中，研究利用乙醇制备 C4 烯烃具有重要的社会现实意义。本文主要针对乙醇耦合制备 C4 烯烃的过程中，催化剂组合以及温度对乙醇转化率、C4 烯烃选择性、C4 烯烃收率的影响进行研究，通过建立基于常见函数拟合算法以及集成学习算法的数学模型，实现了对制备烯烃过程中各影响因素的量化分析，深度刻画了乙醇耦合制备 C4 烯烃问题内在模式。

针对问题一，考虑到需要描绘 24 组数据的变化关系，我们采用简单实用的函数拟合方法，根据数据走势的不同，分别选择线性、二次、指数函数模型对曲线进行拟合，分别刻画了乙醇转化率、C4 烯烃转化率受温度的影响。同时针对某一次特定实验条件下的实验数据，分析了乙醇转化率、各成分选择性对是如何随时间变化的，给出了该实验中控制时间条件的建议。

针对问题二，考虑到催化剂组合的因素都是由文字描述的，我们需要针对其做数据处理，将 Co 负载量、Co/SiO2 与 HAP 的装料比、乙醇浓度、Co/SiO2 的量、HAP 的量这几个变量的数据对应列出。我们首先采用控制变量法进行定性分析，描述乙醇转化率以及 C4 烯烃选择性随各催化剂因素以及温度的变化趋势。接着建立机器学习模型，使用 XGBoost 算法对其进行回归，得到定量拟合结果。

针对问题三，继续沿用前一问的方法，首先使用控制变量法，定性描述 C4 烯烃收率是如何受各催化剂因素以及温度的影响的。接着使用了 AdaBoost、随机森林、XGBoost 这三种集成学习算法进行回归分析，发现在测试集上的打分结果差不多，故再次采用 Voting 集成的思想来对三个回归器的结果进行优化。利用建立好的机器学习模型，寻找 C4 烯烃收率的最大值，我们在此采用了遗传算法来求解，最终得到：在温度为 387.8070℃，Co/SiO2 量为 177.33mg，HAP 量为 178.7992mg，Co 负载量为 1.3383wt%，乙醇浓度为 1.1442 时，C4 烯烃收率达到最大值。当温度限制在 350℃以下时，我们通过数据清洗的手段将 350℃以上的数据丢弃，重新训练上述模型，并再次利用遗传算法进行求解，最终得到：温度为 341.4446℃，Co/SiO2 量为 176.3787mg，HAP 量为 178.2950mg，Co 负载量为 1.9306wt%，乙醇浓度为 1.6519 时，C4 烯烃收率达到最大值。

针对问题四，我们首先使用层次分析法确定了各因素的重要程度，根据重要性排序进行优先考虑，接着尽量弥补原始数据分布不均匀的缺点，使补充后的数据增强可用性，有利于机器学习算法的使用，最后还要考虑了使新补充的数据尽可能与原始数据形成对比实验，这样有利于用控制变量法进行探究。

关键词：函数拟合 控制变量法 集成学习 遗传算法 层次分析法

一、问题重述

C4 烯烃作为生产中的重要化工原料，被广泛应用在生产化工产品和医药中间体中。在传统生产方法中，普遍使用化石能源作为原料，但随之而来的是环境的污染、化石能源的短缺等问题。乙醇作为一种新型的清洁能源，有着来源广泛、绿色环保、可通过生物质发酵制备等优点，以其为平台分子生产高附加值的 C4 烯烃具备巨大的应用前景及经济效益[1]。在 C4 烯烃的制备过程中，Co 负载量、Co/SiO2 和 HAP 装料比、乙醇浓度等催化剂组合以及温度对 C4 烯烃产量的相关数据有重要影响。因此，通过数学建模，对催化剂组合进行设计以提高生产的效能，具有非常重要的意义与价值。

针对问题一，对附件 1 中的每种催化剂组合进行分析，分别就乙醇转化率、C4 烯烃的选择性与温度的关系进行研究，并分析附件 2 中 350 度时给定的催化剂组合在一次实验不同时间的测试结果。

针对问题二，研究不同的催化剂组合及温度对乙醇转化率与 C4 烯烃选择性大小的影响。

针对问题三，给出能使相同实验条件下 C4 烯烃收率尽可能高的催化剂组合与温度的策略。考虑温度低于 350 度时，选择怎样的催化剂组合与温度，能使 C4 烯烃收率尽可能高。

针对问题四，若增加 5 次实验，进一步验证、探究 C4 烯烃收率是如何受催化剂组合以及温度的影响的，给出实验设计及详细理由。

二、问题分析

2.1 问题 1

该问是一个函数拟合的问题，根据题目中所给出的原始数据，我们需要根据催化剂组合不同进行分组，研究乙醇转化率和 C4 烯烃选择性随温度变化的趋势，即它们是如何受温度影响的，并给出显的式函数关系式，确定二者的相关关系。

其次，在某一实验中，我们固定温度和催化剂组合不变，研究该次反应随着时间的延长乙醇转化率的变化和各产物选择性的变化，我们需要观察数据的走势，研究各产物选择性变化是否存在相关关系。

2.2 问题 2

该问研究的是催化剂组合(即 Co 负载量，Co/SiO2 与 HAP 的装料比，乙醇浓度，以及催化剂的量)和温度对乙醇转化率和 C4 烯烃选择性的影响。我们首先定性研究各因素的影响大小，接着使用集成学习的方法来进行模型拟合，得到一个具有确定参数的可量化模型。

2.3 问题 3

该问题在问题 2 的基础上进行，我们已经研究得出了催化剂组合和温度对乙醇转化率和 C4 烯烃选择性的影响，但是这两个模型的变化趋势是不相同的。故 C4 烯烃收率的影响不能简单的将上述两个模型的结果相乘，需要重新构造一个模型，直接的去研究催化剂组合和温度是如何影响 C4 烯烃收率的。在处理完数据后去使用集成学习的算法来进行拟合，得到结果后通过启发式算法来得到该模型的最大值。

2.4 问题 4

由于本题给定的原始数据太少，不足以完全分析出 C4 烯烃收率受催化剂组合以及温度的影响的模式，所以需要重新设计数据，进一步利用控制变量法来研究每种因素的影响。但是由于只能多实验五次，因此按照重要性将影响因素排序，优先针对更加重要的因素来设计实验，进行验证。

三、模型假设

1. 认为该化工实验室实验仪器完好，实验人员操作规范、准确，只存在人力不可避免的偶然误差，其余实验结果均准确无误。
2. 在进行实验时，除了题目中所提到影响因素，即 Co 负载量、Co/SiO2 装料比、乙醇浓度、温度、催化剂的量，其余能够影响实验的因素均保持不变，如压强等。

四、符号说明

表 1

符号名	符号描述	单位
$\alpha_{C_2H_6O}$	C4 烯烃选择性	%
S_{C_4}	C4 烯烃选择性	%
t	反应温度	℃
ω_{Co}	Co 负载量	wt%
η	Co/SiO2 和 HAP 装料比	%
c	乙醇浓度	ml/min

五、模型的求解与建立

5.1 问题一的模型建立与求解

为了建立乙醇转化率以及 C4 烯烃选择性和温度之间关系的数学模型，我们根据给定的数据画出乙醇转化率以及 C4 烯烃选择性随温度变化的折线图，根据图形的大致走势来给出其关系的猜想，并使用 SPSS 统计软件来对曲线进行拟合。

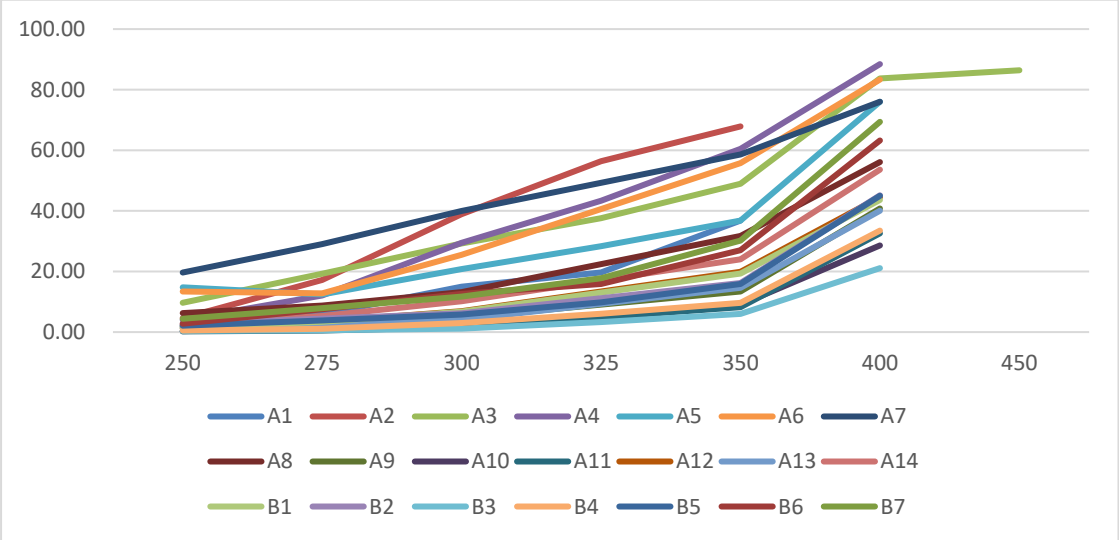


图 1 乙醇转化率随温度变化的情况

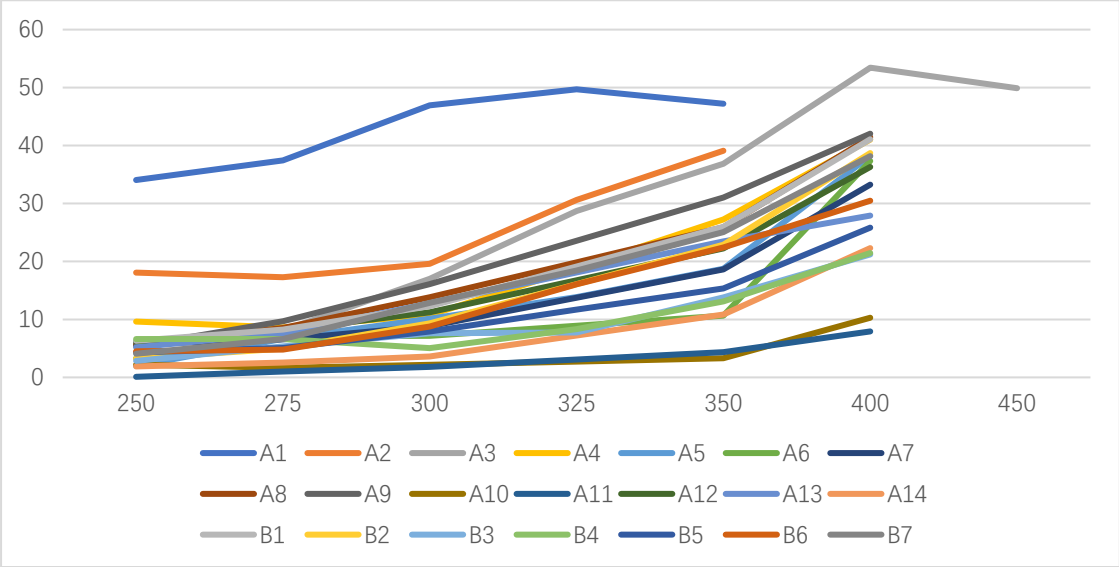


图 2 C4 选择性随温度的变化情况

首先，根据画出来的折线图我们可以看出，基本上所有的催化剂组合情况下，乙醇转化率以及 C4 选择性都是随着温度的上升而增加的，据此我们可以推测出：随着温度的上升，化学平衡正向移动，导致原料的转化率上升；温度的上升导致反应向着多生成 C4 烯烃的方向进行，C4 烯烃选择性上升。

但是，可以发现，在不同的催化剂组合下，上升的速率和幅度不尽相同，故需要我们来建立相关的数学模型来描述这种不同，并针对各种不同的类型来进行相对应的分析。据此，我们提出三种可能的函数关系的猜想，即乙醇转化率以及C4 烯烃选择性随温度变化可能呈现为：一次函数，二次函数，指数函数。

$$\alpha_{C_2H_6O} = \beta_1 t + \beta_0 + \varepsilon \tag{1}$$

$$\alpha_{C_2H_6O} = \beta_2 t^2 + \beta_1 t + \beta_0 + \varepsilon \tag{2}$$

$$\alpha_{C_2H_6O} = \beta_1 e^{\beta_0 t} + \varepsilon \tag{3}$$

其中β_0、β_1、β_2为回归的参数，ε为回归的误差，且$\varepsilon \sim N(0,1)$。

根据计算得到的R^2、F、显著性这三个指标来确定回归拟合得到的曲线是否被采用。如果回归十分显著的话，就确定采用该种回归方式来拟合曲线。如 A1 组使用线性回归确定乙醇转化率与温度的关系曲线为：

$$\alpha_{C_2H_6O} = 0.333t - 84.083 \tag{4}$$

其回归显著性为 0.008，远小于 0.05，即认为两者之间存在着一种线性关系。

接下来我们举例分析每种关系模型，即考虑每种不同的催化剂组合情况下，乙醇转化率以及 C4 烯烃选择性和温度之间的关系。

其中所有的数据统计结果、统计图、拟合结果都存放在附录中。

5.1.1 线性模型

选择 A1、A2、A3、A4 四组的乙醇转化率和温度之间的数据来进行举例。首先将数据的折线图画出：

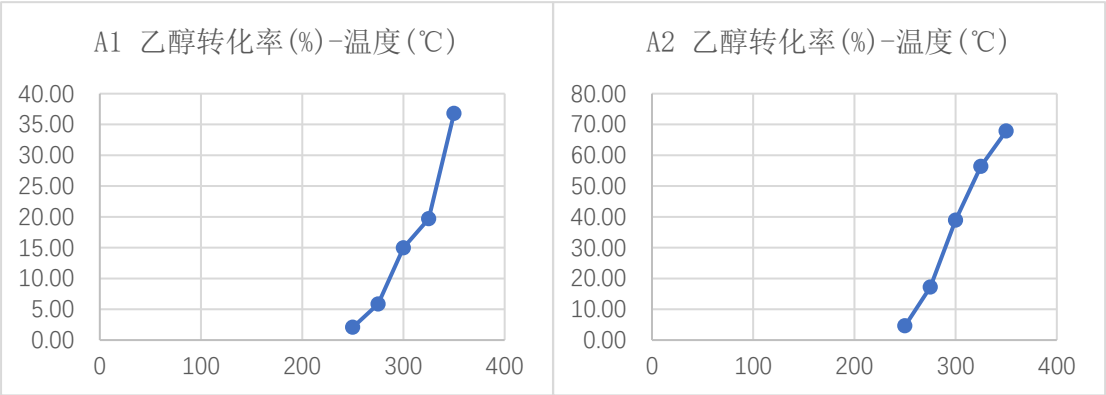


图 3 A1、A2 组乙醇转化率与温度关系图

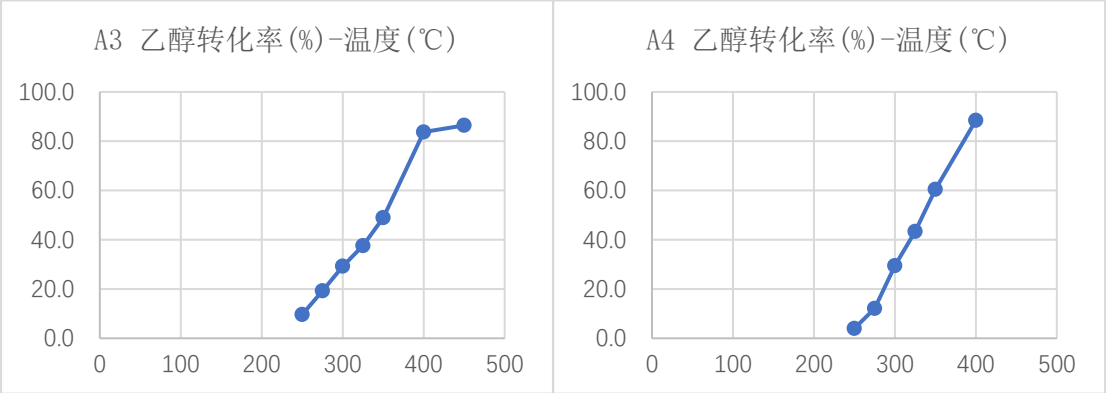


图 4 A3、A4 组乙醇转化率与温度关系图

我们可以从图中直观地发现，乙醇转化率与温度之间存在着一种线性关系，所以我们采用线性函数来进行拟合，得到以下结果：

$$\alpha_{C_2H_6O\_A1} = 0.333t - 84.083 \tag{5}$$

$$\alpha_{C_2H_6O\_A2} = 0.663t - 161.891 \tag{6}$$

$$\alpha_{C_2H_6O\_A3} = 0.42t - 95.883 \tag{7}$$

$$\alpha_{C_2H_6O\_A4} = 0.582t - 144.571 \tag{8}$$

其回归显著性分别为：0.008、0.000、0.000、0.000，故回归十分显著，认为乙醇转化率与温度之间存在线性关系，其中斜率影响了乙醇转化率随温度变化的快慢。（说明：如果显著性为 0.000，这是四舍五入的结果，即说明该数据的最高位数在小数点后 4 位及以后，后面论文的数据如不加说明皆采用该种表示方式）

选择 A1、A3、A8、A9 四组的 C4 烯烃选择性和温度之间的数据来进行举例。首先将数据的折线图画出：

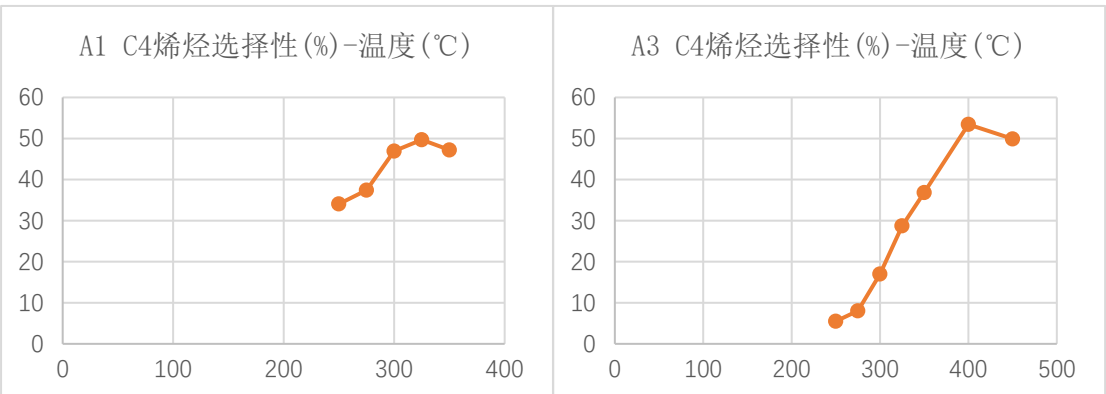


图 5 A1、A3 组 C4 烯烃转化率与温度关系图

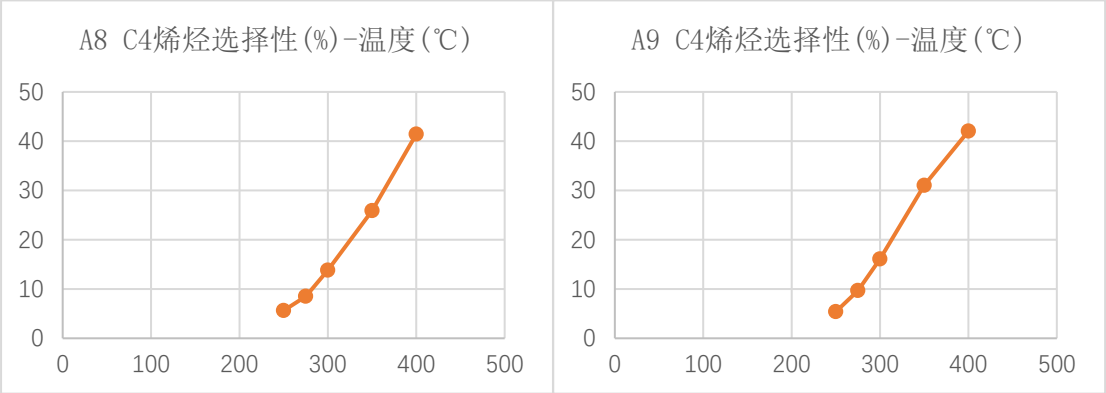


图 6 A8、A9 组 C4 烯烃与温度关系图

我们可以从图中直观地发现，C4 烯烃选择性与温度之间存在着一种线性关系，所以我们采用线性函数来进行拟合，得到以下结果：

$$S_{C_4\_A1} = 0.154t - 3.242 \tag{9}$$

$$S_{C_4\_A3} = 0.261t - 59.195 \tag{10}$$

$$S_{C_4\_A8} = 0.242t - 57.257 \tag{11}$$

$$S_{C_4\_A9} = 0.254t - 59.095 \tag{12}$$

其回归显著性分别为：0.045、0.001、0.001、0.000，故回归十分显著，认为 C4 烯烃选择性与温度之间存在线性关系，其中斜率影响了 C4 烯烃选择性随温度变化的快慢。

说明在温度达到催化剂活性开始降低的点前，乙醇转化率和 C4 烯烃选择性都是随着温度的上升而降低的，如果想要达到较好的生产 C4 烯烃的效率，就要尽量升温。

5.1.2 二次模型

选择 A5、A6 两组的乙醇转化率和温度之间的数据，A10、B4 两组的 C4 烯烃选择性与温度之间的关系来进行举例。首先将数据的折线图画出：

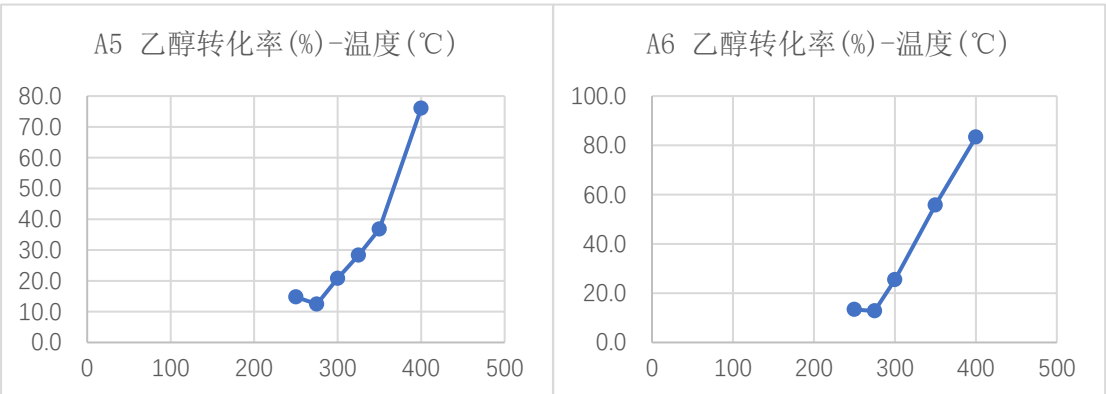


图 7 A5、A6 组乙醇转化率与温度关系图

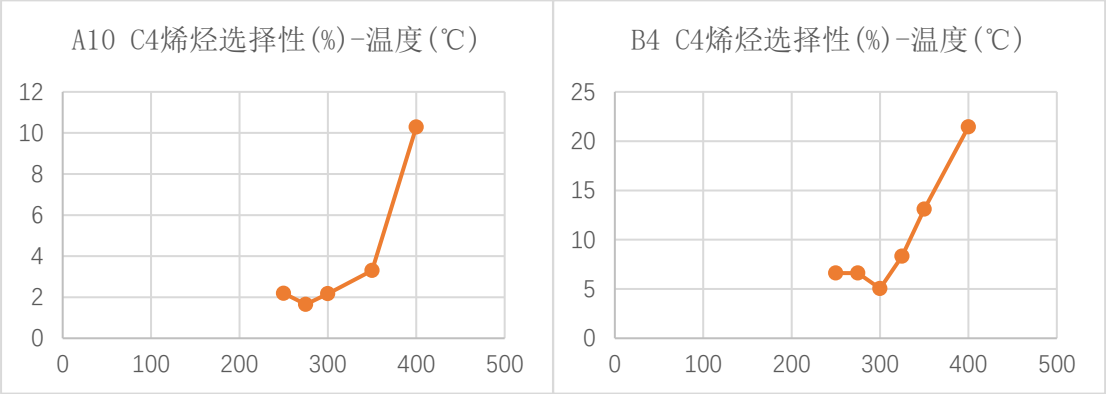


图 8 A10、B4 组 C4 烯烃选择性与温度关系图

根据图中数据表现，我们可以知道这些数据在增长的过程均存在一个“凹陷”，故我们采用二次函数模型来进行拟合，其拟合结果如下：

$$\alpha_{C_2H_6O\_A5} = 0.003t^2 - 1.669t + 231.943 \tag{13}$$

$$\alpha_{C_2H_6O\_A6} = 0.002t^2 - 0.584t + 51.946 \tag{14}$$

$$S_{C_4\_A10} = 0.001t^2 - 0.404t + 59.711 \tag{15}$$

$$S_{C_4\_B4} = 0.001t^2 - 0.549t + 81.074 \tag{16}$$

其回归显著性分别为 0.000、0.014、0.022、0.004，说明回归显著，即乙醇转化率以及 C4 烯烃选择性之间存在一种二次关系，在某个温度点，因变量会陷入某个局部最小值，但是经过升温后，会跳出该最小值点，继续随温度的升高而增大。出现这样的情况可能是由于在某个温度点，催化剂会导致原料朝着某个其他的特定的方向进行转化，且转化率较低，说明该温度并不是十分适合生产 C4 烯烃，需要继续提高温度来寻找效率更加的点。

5.1.3 指数模型

选择 A9、A10、A11、A13 两组的乙醇转化率和温度之间的数据来进行举例。首先将数据的折线图画出：

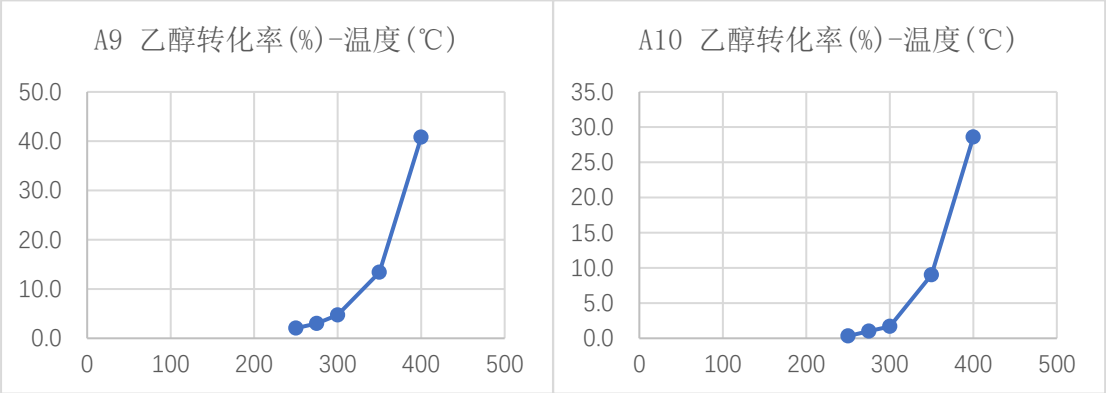


图 9　A9、A10 组乙醇转化率与温度关系图

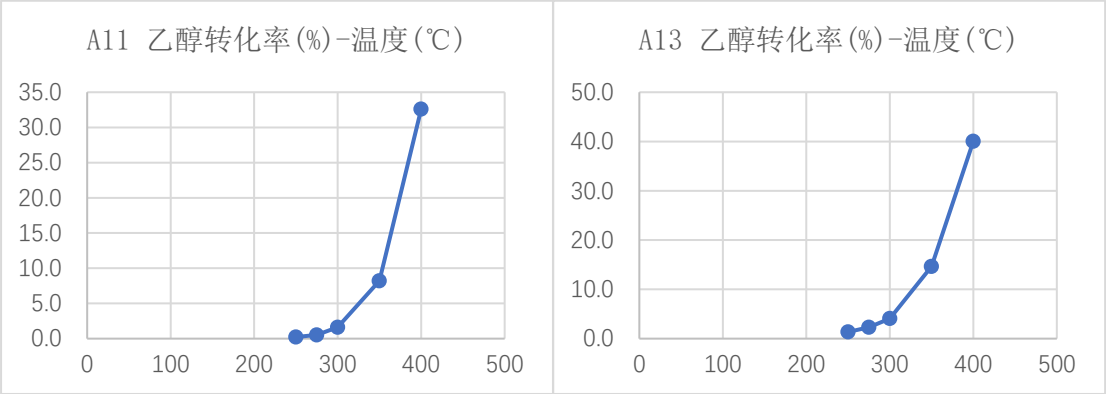


图 10　A11、A13 组乙醇转化率与温度关系图

这一类图有一个明显的特征，即在温度较低的情况下，乙醇的转化率随温度的上升较为缓慢，而在温度再次升高后，乙醇转化率陡增，我们考虑使用指数模型来拟合这样的曲线。拟合结果如下：

$$\alpha_{C_2H_6O\_A9} = 0.012e^{0.02t} \tag{17}$$

$$\alpha_{C_2H_6O\_A10} = 2.49 \times 10^{-4} \cdot e^{0.029t} \tag{18}$$

$$\alpha_{C_2H_6O\_A11} = 5.778 \times 10^{-5} \cdot e^{0.033t} \tag{19}$$

$$\alpha_{C_2H_6O\_A13} = 0.004 \cdot e^{0.023t} \tag{20}$$

出现该种情况的原因可能是，在温度并不太高的情况下，反应并没有被完全激活，乙醇转化率随温度的上升保持平稳或略微增加，但是当到达或超过该反应的一个温度阈值后，反应会被激活，进入另一种状态，乙醇转化率随着温度的上升幅度大幅度提升。

5.1.4 乙醇转化率以及 C4 烯烃选择性与温度的量化关系

表 2 乙醇转化率以及 C4 烯烃选择性与温度的量化关系表

组别	乙醇转化率与温度的关系	C4 烯烃选择性与温度的关系
A1	$\alpha_{C_2H_6O} = 0.333t - 84.083$	$S_{C_4} = 0.154t - 3.242$
A2	$\alpha_{C_2H_6O} = 0.663t - 161.891$	$S_{C_4} = 0.222t - 41.546$
A3	$\alpha_{C_2H_6O} = 0.42t - 95.883$	$S_{C_4} = 0.261t - 59.195$
A4	$\alpha_{C_2H_6O} = 0.582t - 144.571$	$S_{C_4} = 0.227t - 52.411$
A5	$\alpha_{C_2H_6O} = 0.003t^2 - 1.669t + 231.943$	$S_{C_4} = 0.23t - 57.813$
A6	$\alpha_{C_2H_6O} = 0.002t^2 - 0.584t + 51.946$	$S_{C_4} = 0.203t - 50.748$
A7	$\alpha_{C_2H_6O} = 0.378t - 74.260$	$S_{C_4} = 0.187t - 44.262$
A8	$\alpha_{C_2H_6O} = 0.002t^2 - 0.801t + 96.863$	$S_{C_4} = 0.242t - 57.257$
A9	$\alpha_{C_2H_6O} = 0.012e^{0.02t}$	$S_{C_4} = 0.254t - 59.095$
A10	$\alpha_{C_2H_6O} = 2.49 \times 10^{-4} \cdot e^{0.029t}$	$S_{C_4} = 0.001t^2 - 0.404t + 59.711$
A11	$\alpha_{C_2H_6O} = 5.778 \times 10^{-5} \cdot e^{0.033t}$	$S_{C_4} = 0.052t - 13.307$
A12	$\alpha_{C_2H_6O} = 0.002t^2 - 0.954t + 121.526$	$S_{C_4} = 0.205t - 47.727$
A13	$\alpha_{C_2H_6O} = 0.004 \cdot e^{0.023t}$	$S_{C_4} = 0.163t - 35.889$
A14	$\alpha_{C_2H_6O} = 0.002t^2 - 1.083t + 137.991$	$S_{C_4} = 0.138t - 35.116$
B1	$\alpha_{C_2H_6O} = 0.002t^2 - 0.948t + 121.238$	$S_{C_4} = 0.24t - 56.728$
B2	$\alpha_{C_2H_6O} = 0.026 \cdot e^{0.019t}$	$S_{C_4} = 0.244t - 61.073$
B3	$\alpha_{C_2H_6O} = 3.45 \times 10^{-4} \cdot e^{0.028t}$	$S_{C_4} = 0.12t - 28.294$
B4	$\alpha_{C_2H_6O} = 0.001 \cdot e^{0.027t}$	$S_{C_4} = 0.001t^2 - 0.549t + 81.074$
B5	$\alpha_{C_2H_6O} = 0.014 \cdot e^{0.02t}$	$S_{C_4} = 0.146t - 34.599$
B6	$\alpha_{C_2H_6O} = 0.003t^2 - 1.445t + 190.083$	$S_{C_4} = 0.19t - 45.757$
B7	$\alpha_{C_2H_6O} = 0.05 \cdot e^{0.018t}$	$S_{C_4} = 0.234t - 56.451$

所有回归结果均经过回归检验，回归显著，具体的检验量化结果见附件。

5.1.5 350 度时给定的催化剂组合结果分析

根据我们的数据可知，在 350℃，给定某种催化剂的情况下，随着时间的增加，乙醇的转化率在逐渐下降，且较为明显。可能是由于在该温度下，反应的其他产物分解，微粒堵塞了催化剂的活性位，导致了催化剂活性的下降，使得反应的转化率下降。

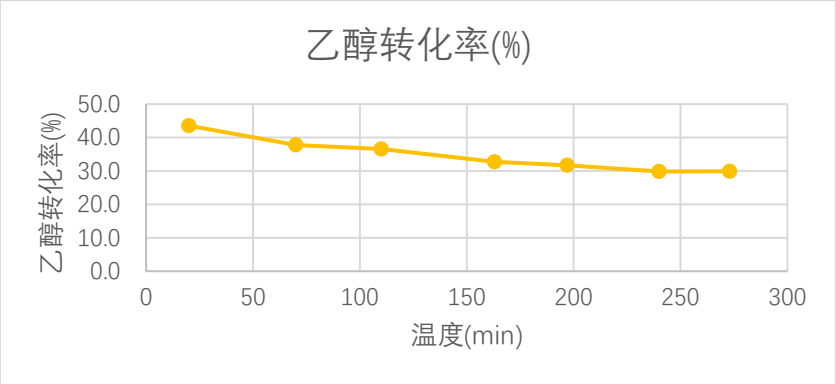


图 11 乙醇转化率随温度的变化

同时我们可以看到，各种产物的选择性随着时间的变化情况：

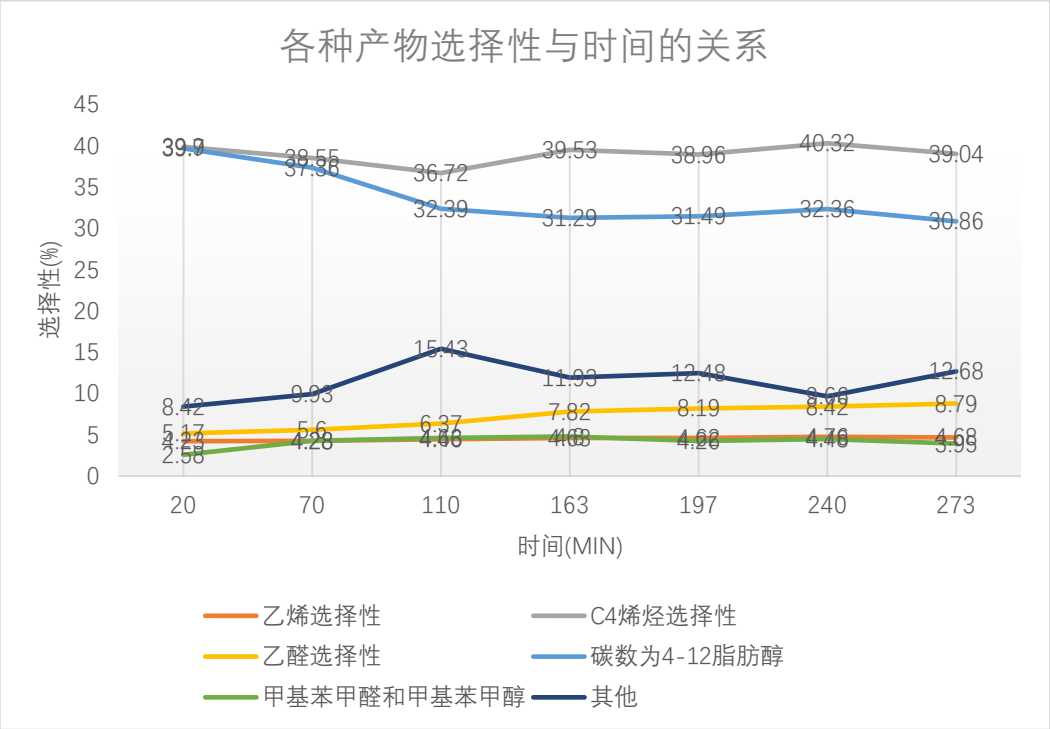


图 12 各种产物选择性随时间的变化关系

可以看到在 20min 到 110min 之间，C4 烯烃选择性与碳数为 4-12 脂肪醇选择性略有下降，且其他成分的选择性略有上升，可能是因为在该段时间，C4 烯烃和 4-12 脂肪醇可能由于某些原因分解了，导致了其他成分的增加，并且由于这些其他成分中的微粒堵塞了催化剂，导致了乙醇转化率的下降。

我们研究的目的是探究制备 C4 烯烃的工艺条件，也就是说我们主要关心的 C4 烯烃的选择性。

综合整段时间，即从 20min 到 273min，我们可以发现 C4 烯烃的选择性基本是稳定的，也就是说只要该反应还在进行，产物中 C4 烯烃的含量就是稳定的。所以影响 C4 烯烃生产的主要就是乙醇的转化率，而随着实验时间的延长，我们可以发现乙醇转化率显著下降，也就是说该反应只能在一段时间内达到最佳制备效果。过了该段时间后，生产效率会大幅下降，此时需要分离所需的 C4 烯烃，清理其余不需要的产物，并即时更换催化剂，以保证 C4 烯烃制备的高效性。

5.2 问题二的模型建立与求解

首先明确我们要探究的是关系的自变量是催化剂组合方式和温度，但是数据中体现出来的能够影响乙醇转化率和 C4 烯烃选择性的因素还有一个：即装料方式，有 A、B 两种。控制其他因素不变，分别画出乙醇转化率以及 C4 烯烃选择性在装料方式不同的情况下的图，可以发现两者相差其实并不大。其次，通过方差分析，可以明确的确定两者的 F 值很小，显著性很大，即可以认为装料方式对

乙醇转化率和 C4 烯烃选择性的影响很小，在以下的研究种认为其对乙醇转化率和 C4 烯烃选择性没有影响。

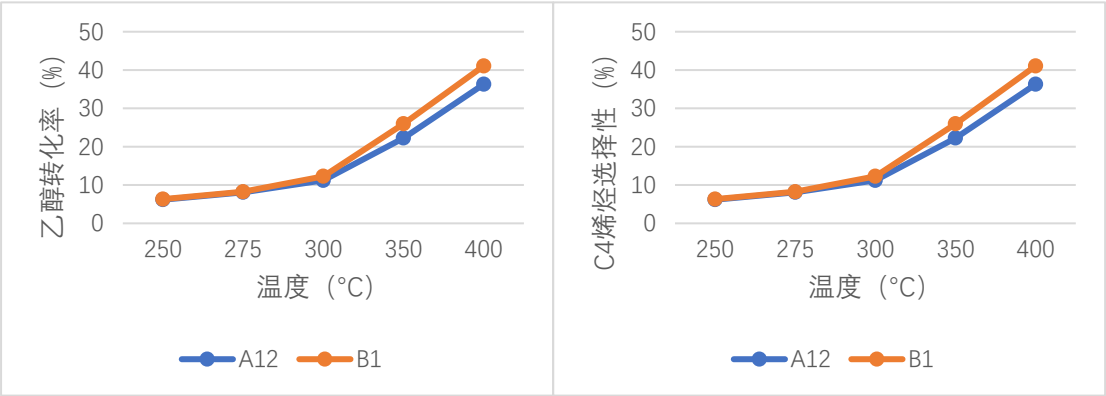


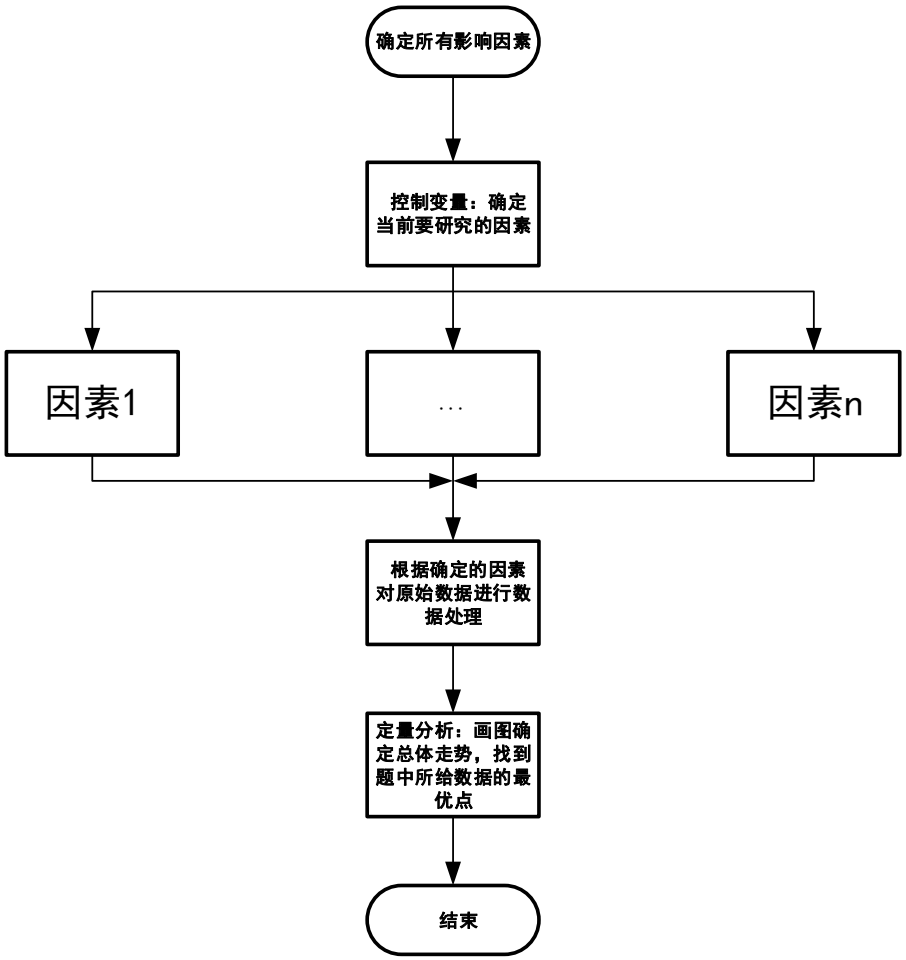
图 13 A12、B1 乙醇浓度于 C4 烯烃选择性随温度的变化关系

表 3 方差检验结果

	F	显著性
乙醇转化率	0.001	0.975
C4 烯烃选择性	0.052	0.825

5.2.1 研究方法

在研究乙醇转化率以及 C4 烯烃选择性是如何受催化剂组合以及温度的影响时，我们需要使用控制变量法，即欲确定某一因素的影响时，需要控制其他的因素保持不变，只改变该因素，从而得到观察结果，并进一步分析。在得到定性分析结果后，我们仍然需要尝试使用数学模型来得到定量拟合的结果。



我们要研究的影响乙醇转化率和 C4 烯烃选择性的催化剂因素有以下几种：

表 4 影响因素说明

因素名称	因素说明	因素单位
Co 负载量	$M(Co)/SiO2$	1wt%
Co/SiO2 和 HAP 装料比	$M(Co/SiO2)/HAP$	%
乙醇浓度	每分钟加入乙醇的毫升数	ml/min
温度	反应温度	℃

5.2.2 定性分析

首先研究 Co 负载量的影响：

通过题目给的数据可以观察到，在 A1、A2、A4、A6 组，只有 Co 的负载率在发生变化，而 Co 负载量保持为 1:1，乙醇浓度保持为 1.68ml/min。

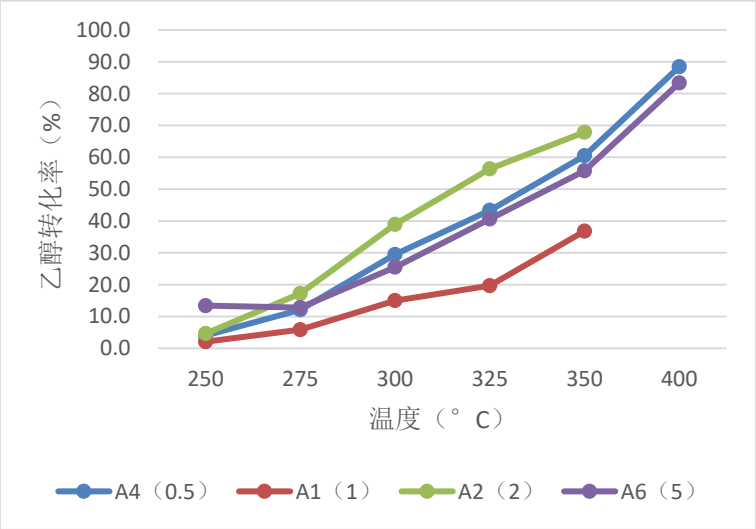


图 14　A4、A1、A2、A6 组乙醇转化率随温度的变化

通过上过可以发现在 275℃到 325℃之间，乙醇转化率在 Co 负载量为 2wt%时最好，而在 400℃时候，由于缺失数据，无法确定 2wt%是否仍然为最优条件，但是可以大致上肯定 2wt%仍然为较优数据。

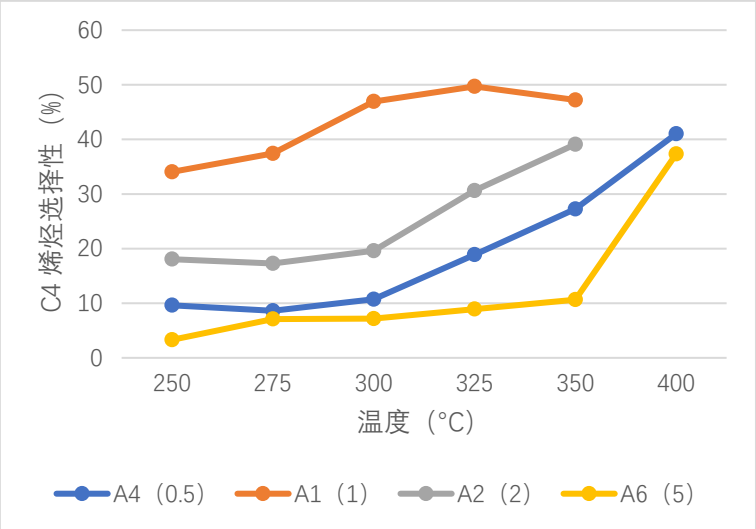


图 15　A4、A1、A2、A6 组 C4 烯烃选择性随温度的变化

通过在 350℃之前 Co 负载量为 1wt%使得 C4 烯烃选择性一直具有显著的优势，推断其在 400℃时仍然具有较优表现。

接着研究 Co/SiO2 和 HAP 装料比的影响：

原始数据中有一组特殊的数据，它是唯一没有装载 HAP 的情况，即 A11 组，同时发现 A12 组与其仅在 HAP 这一点反应条件不同，其余均保持 Co 负载量为 1wt%，乙醇浓度为 $1.68ml/min$。

可以发现在所有温度下，A12 的表现情况均优于 A11。所以可以认为装载 HAP 对该反应极其重要，对提升乙醇转化率以及 C4 烯烃选择性具有确定的效果。确定了一定要装载 HAP，那么到底要装载多少 HAP 呢？

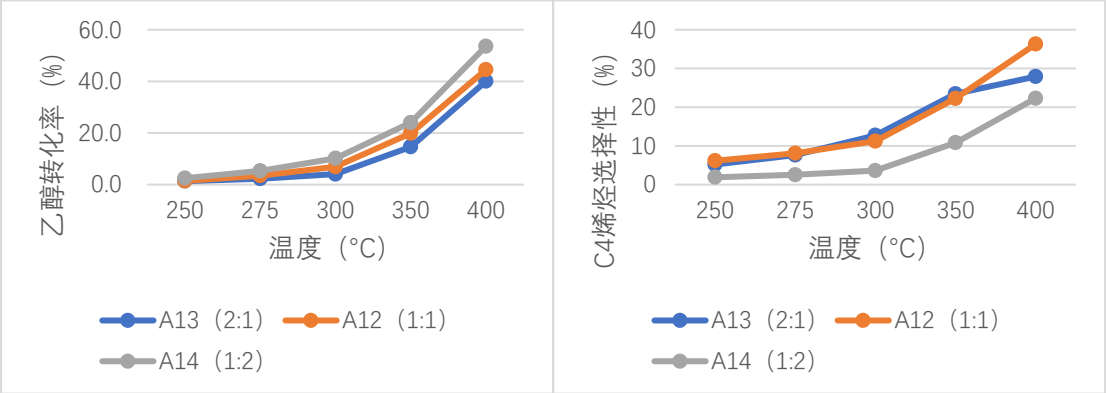


图 16　A12、A13、A14 组乙醇转化率以及 C4 烯烃选择性对温度的变化

在图中可以看到，装料比对乙醇转化率和 C4 烯烃选择性的影响不尽相同，需要分开考虑：

在不同的温度下，我们可以观察到，装料比分别为 2:1、1:1、1:2 时，乙醇转化率呈现递增关系，但是差别并不十分大，说明装料比对乙醇转化率只有微小的影响。

而对于 C4 烯烃选择性来说，装料比为 1:2 时表现最差，在 350℃之前，装料比为 1:1 和 2:1 的表现差不多，但是在温度为 400℃时，装料比为 1:1 的 C4 选择性显著的优于装料比为 2:1 和 1:2 的情况。也就是说，在温度为 400℃时存在这样一个明显的规律：随着 Co/SiO2 和 HAP 装料比的上升，C4 烯烃选择性先变大后变小。

其次研究乙醇浓度的影响：

乙醇是反应物，反应物的浓度在化学反应是一个十分重要的分析因素，根据化学常识，反应物的浓度越高，反应物就会与催化剂接触的越充分，从而反应的就会越充分。我们通过 A7、A8、A9 三组数据来探究乙醇浓度对乙醇转化率以及 C4 选择性的影响，这三组数据的 Co 负载量保持为 1wt%，Co/SiO2 和 HAP 装料比保持为 1:1。

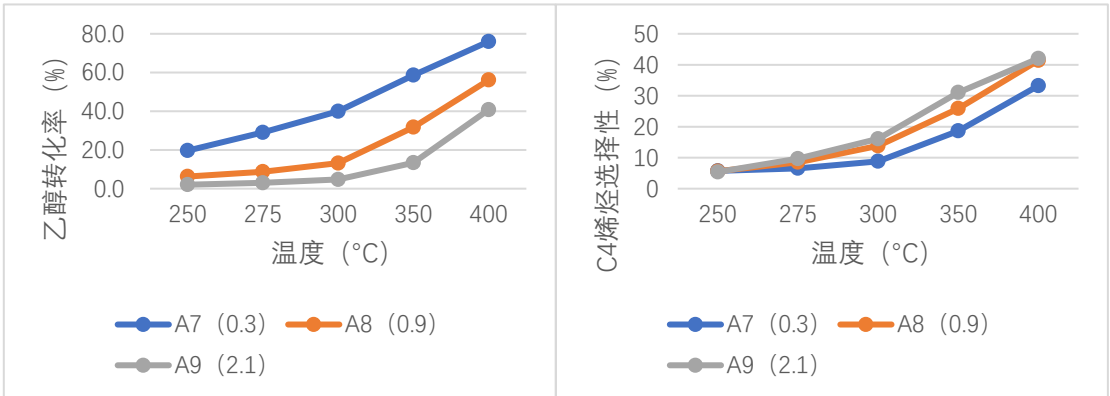


图 17　A7、A8、A9 组乙醇转化率以及 C4 烯烃选择性对温度的变化

可以看出随着乙醇浓度的上升，乙醇转化率不断减小，且规律十分明显。

随着乙醇浓度的上升，C4 烯烃选择性不断增大，但是在 400℃时乙醇浓度为 0.9 和乙醇为 2.1 的结果几乎相同，猜测可能是由于从 0.9 到 2.1 跨度较大，历经 C4 选择性先增大后减小的过程。并且由于在不同乙醇浓度下的结果相差并不十

分多，所以可认为乙醇浓度对 C4 烯烃选择性的影响并不大。

紧接着研究催化剂量的影响：

虽然题目中提到的催化剂组合只有三个因素，即 Co 负载量、Co/SiO2 和 HAP 装料比、乙醇浓度，但是我们通过观察数据还可以发现一个对乙醇转换率和 C4 选择性影响较为明显的因素，即催化剂的量。

我们沿用控制变量的方法，保持 Co 负载量为 1wt%，Co/SiO2 和 HAP 装料比为 1:1，乙醇浓度为 $1.68ml/min$。由于装料比为 1:1，故下面研究时催化剂的量都指的是 Co/SiO2 的量。

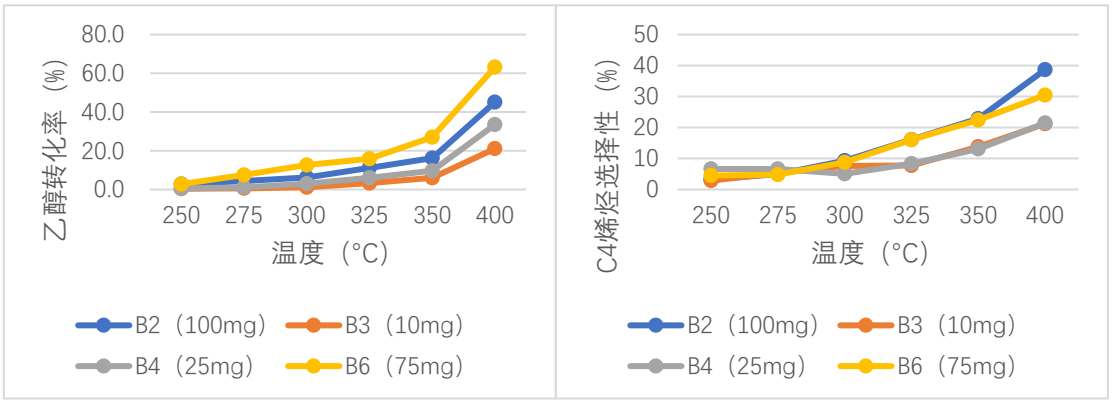


图 18　B2、B3、B4、B6 组乙醇转化率以及 C4 烯烃选择性对温度的变化

可以发现，乙醇转化率在个温度下均随着催化剂量的增大先增大后减小。从 10mg 到 75mg 增大，之后从 75mg 到 100mg 减少。

而 C4 选择性的变化规律比较特别，10mg 和 25mg 在个温度下表现情况基本相同。在 300℃之后，75mg 和 100mg 的 C4 选择性显著高于 10mg 和 25mg 的情况，同时在 400℃时，100mg 的情况优于 75mg 的情况。

最后研究温度的影响：

根据 5.1 中对温度和乙醇转化率以及 C4 烯烃选择性大小的分析可以知道，随着温度的上升，二者皆呈现递增的趋势，不同的是受催化剂组合影响而导致的增长趋势与速率的不同。

5.2.3 定量分析 C4 烯烃收率随各催化剂因素和温度的影响

通过数据可以发现，在 Co/SiO2 装料比这个因素下，只有一组数据为 1:2，一组数据为 2:1，其余所有的数据均为 1:1，也就是这一因素的影响变化由于数据的原因，我们几乎不能用定量分析的方式来确定一个模型。

故我们将影响因素做一个转换，从研究 Co/SiO2 装料比、催化剂的量这两个因素对乙醇转化率、C4 烯烃选择性的影响转换为研究 Co/SiO2 的量、HAP 的量这两个因素对 C4 烯烃收率的影响。即总的来说，我们的研究目标是：探究 Co 负载量、乙醇浓度、温度、Co/SiO2 的量、HAP 的量这五个因素对对乙醇转化率、C4 烯烃选择性的影响。

在开始模型设计之前，首先要对数据进行处理，从对催化剂组合的文字性描述一栏中，我们使用了 EXCEL 软件分别抽取出了 Co/SiO2 的量、HAP 的量这两项因素的量化水平。同时这样处理后的数据仍然会被应用到第三问。

根据前面的分析，我们需要拟合这样两个模型：

表 5　待拟合模型的因变量与自变量

因变量	自变量
乙醇转化率	Co 负载量 乙醇浓度 温度 Co/SiO2 的量 HAP 的量
C4 烯烃选择性	Co 负载量 乙醇浓度 温度 Co/SiO2 的量 HAP 的量

这是一个多元回归问题，但是我们发现使用多元线性回归的效果并不好，故使用了 XGBoost 算法来进行拟合。

XGBoost 算法的原理：Boosting 的基本思想是通过某种方式使得每一轮基学习器在训练过程中更加关注上一轮学习错误的样本，不同 Boosting 算法的区别在于是采用何种方式。而 Gradient Boosting 中将负梯度作为上一轮基学习器犯错的衡量指标，在下一轮学习中通过拟合负梯度来纠正上一轮犯的错误。更进一步，XGBoost 就是在 Gradient Boosting 原损失函数的基础上添加了正则化项产生了新的目标函数。二是对目标函数进行二阶泰勒展开，以类似牛顿法的方式来进行优化。[2]

我们将数据切分为训练集和测试集，使用训练集训练参数，通过测试集及逆行测试，乙醇转化率、C4 烯烃选择性的模型测试分数分别为。说明拟合结果较优。

5.3 问题三的模型建立与分析

在问题二中已经建立了乙醇转化率以及 C4 烯烃选择性受催化剂组合以及温度的影响的数学模型，而此时我们需要探究 C4 烯烃收率受催化剂组合以及温度的影响，C4 烯烃收率定义如下：

$$\text{C4烯烃收率} = \text{乙醇转化率} \times \text{C4烯烃选择性}$$

虽然定义 C4 烯烃收率的两个因素在上一问均已分别讨论过了，但是可以明显地观察到，催化剂组合以及温度对这两个因素的影响的趋势并不相同，因此需要从整体上对其进行讨论。

首先进行数据处理，根据每个组别的催化剂组合的描述，分别提取出 Co 的负载量、Co/SiO2 和 HAP 装料比、乙醇浓度、催化剂的量这几个指标，并分别对应列入表中，计算每组数据的 C4 烯烃收率。

为了得到一个直观的印象，画出 24 组数据 C4 烯烃收率变化的曲线。

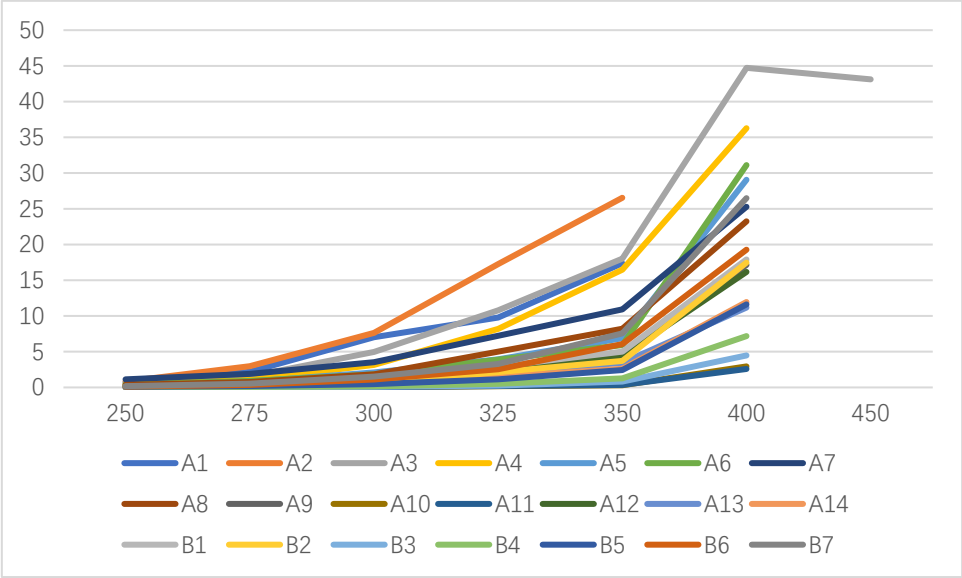


图 19 C4 烯烃收率随温度的变化

接着，沿用上一问的方法，我们采用控制变量法确定每种因素，即 Co 负载量、Co/SiO2 和 HAP 装料比、乙醇浓度、催化剂的量、温度这五个指标，是如何影响 C4 烯烃收率的。

分别讨论完各个因素对 C4 烯烃收率的影响后，如果选择每个因素下的最优水平进行组合，得到的 C4 烯烃收率可能并不是一个较好的值，因为本模型并没有综合分析，并不知道这几个催化剂组合因素和温度之间会不会有什么相互间的影响，故我们还应将所有的影响因素综合起来进行分析。

对于本模型，首先从定性的角度去分析，找到一个较优的组合，但是只从定性的角度考虑是不可靠的，我们还需要根据数据设计某种能够进行模式识别的模型，对于本题，主要考虑使用机器学习的方式对模型进行定量拟合。

5.3.1 定性分析

首先研究的是 Co 的负载量的影响：

通过控制 Co/SiO2 和 HAP 装料比、乙醇浓度、催化剂的量这三个指标保持为 1:1、$1.68ml/min$、$400mg$，对比各组温度下的 C4 烯烃收率。

可以发现 350℃之前，在 Co 负载量为 *2wt%*时，C4 烯烃收率显著高于其他 Co 负载量下的水平。即可以认为，Co 的负载量在 *2wt%*时具有较优结果。并且可以发现，在 Co 负载量为 5wt%时，其 C4 烯烃收率反而较小，说明一味地增大 Co 负载量并不能提高 C4 烯烃收率，反而可能会改变化学反应的环境，致使反应向其他方向进行。但是由于缺乏 400℃的数据，并不好说 *2wt%*仍为较优 Co 负载量。

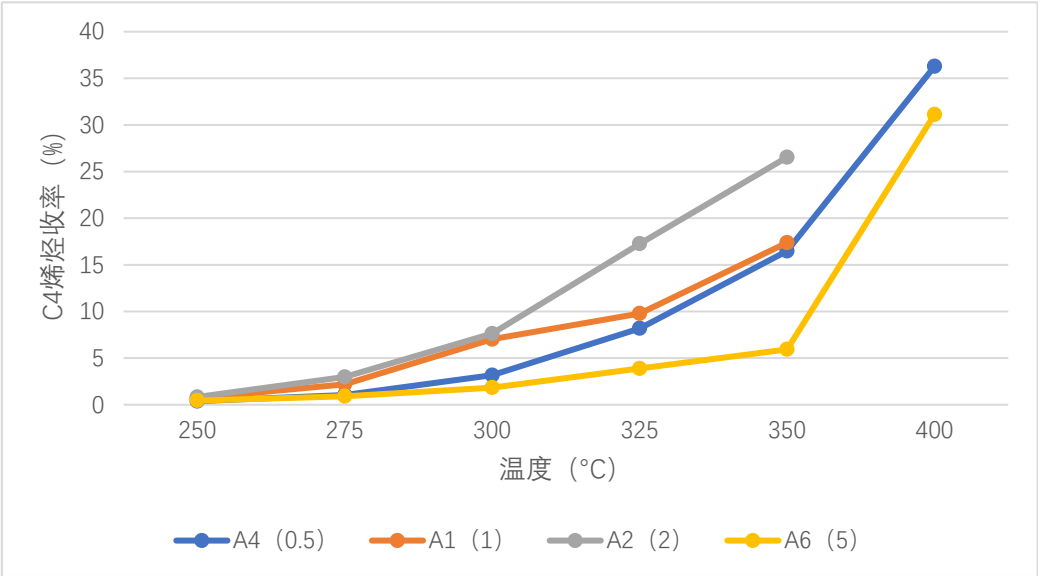


图 20　A4、A1、A2、A6 组 C4 烯烃收率随温度的变化

接着研究 Co/SiO2 和 HAP 装料比的影响：

我们控制 Co 负载量、乙醇浓度、催化剂的量这三个指标保持为 1$wt\%$、1.68ml/min、100mg，对比各组温度下的 C4 烯烃收率。

据图可以看出，Co/SiO2 和 HAP 装料比为 1:1 时，C4 烯烃收率的值显著优于其他组，说明 1:2 和 2:1 组的比例会导致反应的改变，Co/SiO2 和 HAP 装料比是一个中间型指标，当量保持在一个中间水平时，会导致 C4 烯烃收率保持一个较优结果。从化学机理上看，可能是 Co 和 HAP 分别会导致反应向一对相反的方向发展，只有将二者保持在一个较为均衡的水平，才能保证 C4 烯烃收率的水平。

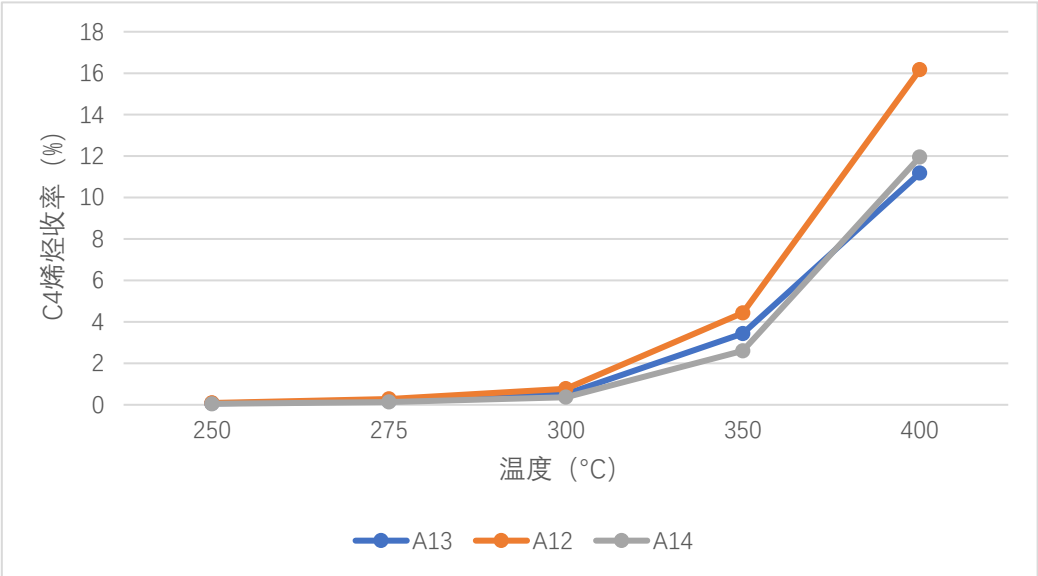


图 21　A12、A13、A14 组 C4 烯烃收率随温度的变化

其次，研究乙醇浓度的影响：

我们控制 Co 负载量、Co/SiO2 和 HAP 装料比、催化剂的量这三个指标保

持为 1$wt\%$、1:1、100mg，对比各组温度下的 C4 烯烃收率。

可以发现，这三条曲线的走势十分一致，且在乙醇浓度为 0.3 时，C4 烯烃收率较优。根据上一问的分析可知，随着乙醇浓度的上升，C4 烯烃选择性会上升，即乙醇浓度的上升会引导反应朝着多生成 C4 烯烃的方向进行，但是根据数据可知，这种引导强度并不大，只是带来少量的 C4 烯烃选择性的提升；相反，乙醇浓度的上升会较明显地抑制乙醇转化率，这符合化学常识，即勒夏特列原理，化学平衡会朝着减弱浓度上升的方向移动。

综合这两个因素，由于乙醇浓度的上升对乙醇转化率的影响比对 C4 烯烃选择性的影响更加明显，故总体上来说，乙醇浓度的上升会导致 C4 烯烃收率的下降。

即认为乙醇浓度为 0.3ml/min时，C4 烯烃收率较优。

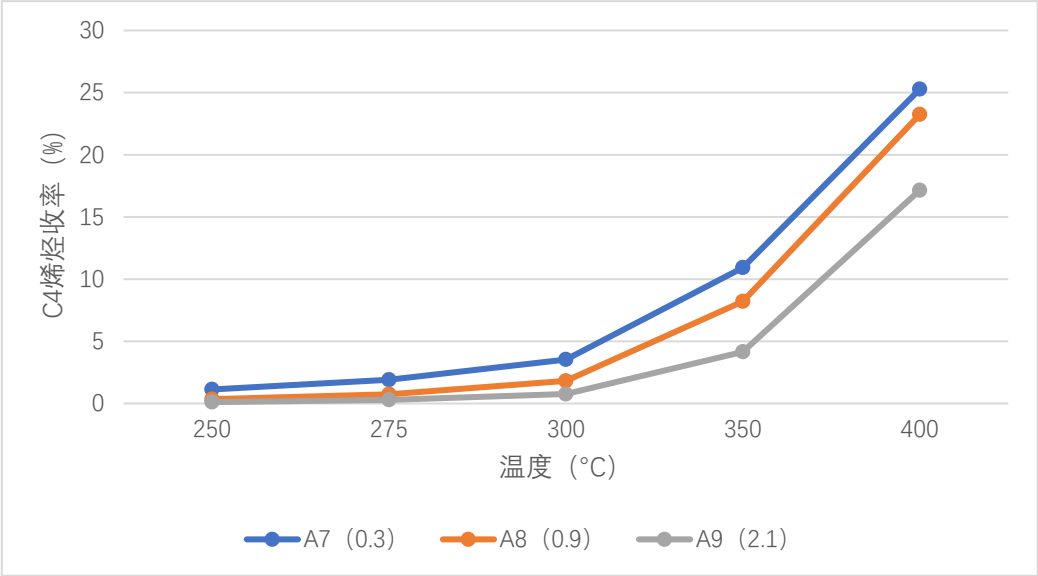


图 22　A7、A8、A9 组 C4 烯烃收率随温度的变化

紧接着研究催化剂量的影响：

我们控制 Co 负载量、Co/SiO2 和 HAP 装料比、乙醇浓度这三个指标保持为 1$wt\%$、1:1、1.68ml/min，对比各组温度下的 C4 烯烃收率。由于装料比为 1:1，故我们在描述催化剂的量时只以其中的一个部分为代表，即 200mg 代表 Co/SiO2 为 200mg，HAP 也为 200mg。

可以发现，在 350℃之前，200mg 催化剂量水平下的 C4 烯烃收率一直远高于其他催化剂量水平下的 C4 烯烃收率，故我们有理由认为，在 350℃之前，200mg 为较优催化剂量。但是在 350℃时候，我们缺乏 200mg 催化剂量的 C4 烯烃收率数据，故暂且先不讨论 200mg 催化剂水平。在这个温度下，50mg、75mg、100mg 催化剂量的 C4 烯烃收率水平差不多，75mg 略大，故认为在 400℃左右时，在 75~100mg 这个区间内，C4 烯烃收率的水平基本保持不变，且可能呈现先增大再减小的趋势，如果在超过 100mg 这个水平的催化剂量时，C4 烯烃收率没有显著增加，则认为 75mg 为催化剂量的一个较优因素。

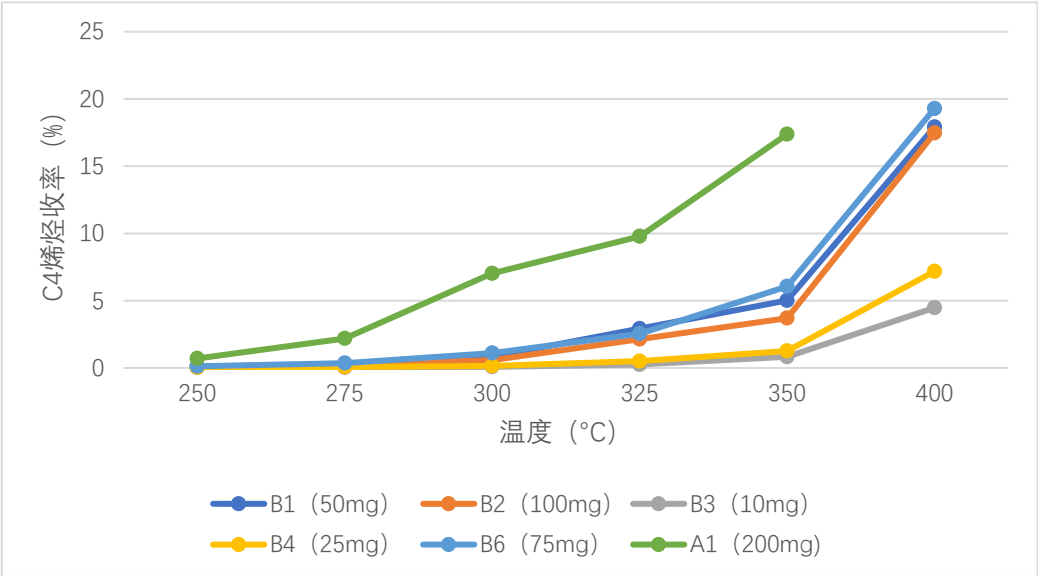


图 23　B1、B2、B3、B4、B6、A1 组 C4 烯烃收率随温度的变化

最后研究温度的影响：

通过前面几个因子的作图分析，我们可以发现，在 400℃之前，无论在什么样的催化剂组合下，温度都是越高越好的，即：随着温度地升高，C4 烯烃收率也在增加，二者成正相关。但是我们由 A3 组的数据可知，在 400℃之后，C4 烯烃收率可能会随着温度的升高而下降。

由于只是简单的定性讨论，故我们只在已经给出的可能中选择一个较优的即果。综合以上因素，我们有一定的理由认为：选择 200mg 1wt%Co/SiO2- 200mg HAP-乙醇浓度 0.9ml/min 的催化剂组合、400℃的反应温度，会是 C4 烯烃收率达到一个较优的水平。

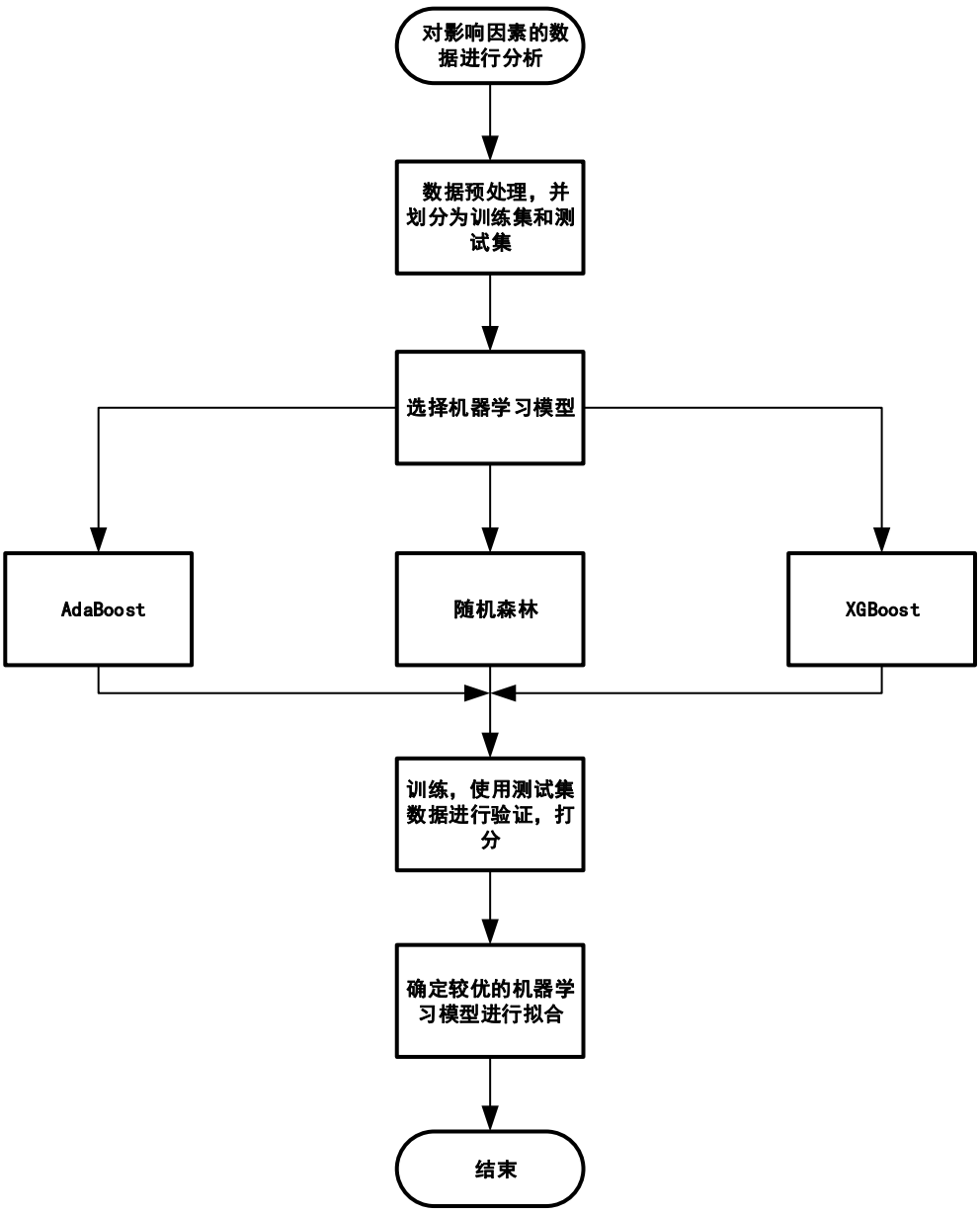
5.3.4 集成学习拟合 C4 烯烃收率与催化剂组合以及温度的关系

在本问前面的分析中，我们都是通过控制变量法，将每个因素的各个水平看成是是离散的，进而讨论某种可能的催化剂因素和温度的组合。而在此处，我们为了综合考虑所有因素对 C4 烯烃收率的影响，使用了机器学习的方法对整个数学模型进行了拟合。

之前的讨论中，我们得出这样的结论：影响 C4 烯烃收率的因素共有五个，分别是 Co 负载量、Co/SiO2 装料比、乙醇浓度、温度、催化剂的量，他们对 C4 烯烃收率影响的方式各不相同。

由第二问的分析，我们将以上五个因素转化为 Co 负载量、乙醇浓度，温度，Co/SiO2 的量、HAP 的量，并研究它们是如何影响 C4 烯烃收率的。沿用第二问已经处理过的数据。

整个研究流程如下图所示。



我们确定的数据划分比例为训练集:测试集=9:1，即将原始数据导入后，做 shuffle 处理，接着按照 9:1 的数量比，把原始数据集划分为训练集和测试集。

我们首先排除了使用线性模型进行拟合的可能，因为从图中可以很明显地发现并不是所有的因素对 C4 烯烃收率的都是线性的。因此我们初步选择了三种模型，分别是 AdaBoost，随机森林，XGBoost，我们分别对在这三种模型的情况下进行基本的调参，确定一组较优的超参数，由于超参数的微小变化，比如基估计器的数目(在随机森林模型中即为决策数的数目)，对模型的效果影响并不大，我们只要确定一个效果较优的范围，并不用确定最精确的超参数。

这三个模型原理如下：

表 6 三种集成学习算法原理

模型名称	模型原理
AdaBoost	Adaboost 是一种迭代算法，其核心思想是针对同一个训练集训练不同的分

	类器(弱分类器)，然后把这些弱分类器集合起来，构成一个更强的最终分类器，是一种集成学习
RandomForest	随机森林是一个包含多个决策树的分类器，并且其输出的类别是由个别树输出的类别的众数而定，是一种集成学习
XGBoost	是 GradientBoosting 算法的一种提升，采用梯度下降的方式修正每一轮学习的错误，进而提升算法的准确性，是一种集成学习

在调完每个模型的参数后，我们重新使用训练集训练一遍，确定每个模型中的所有具体参数，并使用测试集进行评估与打分，打分结果越靠近 1 说明拟合的效果越好。

表 7 三种集成学习算法在测试集上的得分

模型名称	测试集上的得分
AdaBoost	0.971202836444328
RandomForest	0.9867438592369263
XGBoost	0.9902072080978709

从分数上可以看出，在测试集中，这三个模型的表现差不多，并且这三种算法本质上都是集成学习，我们使用投票的方式来确定我们最终的模型，即再次应用采用集成学习 Voting 的方法。

训练好的 VotingRegressor 模型保存在支撑材料中。

我们当前已经找到一个能够从量上描述 C4 烯烃收率随几个催化剂因素和温度变化的模型，接下来的任务就是根据给定的模型找到对应的最大值，首先我们要确定每个参数的限定条件，我们将每个因素的上下界分别设定为题目中所给原始数据的最大值和最小值，即：

表 8 各因素的搜索上下界

因素名	上界	下界	单位
Co 负载量	0.5	5	*wt%*
乙醇浓度	0.3	2.1	*ml/min*
温度	250	400	℃
Co/SiO2 的量	10	200	*mg*
HAP 的量	10	200	*mg*

那么如何确定最优值呢，我们首先尝试使用遍历搜索的办法进行寻找，发现如果想要得到一个较精确的结果，就需要把每一个因素的的迭代步长设置的较小，如温度可能需要将步长设置为 0.5℃，遍历搜索的时间复杂性极其的高。再加上被搜索的模型是一个集成模型，根据输入得到输出本身就需要比初等函数长的多的时间，故使用遍历搜索在有限时间内得到结果是不现实的。

我们采用遗传算法来解决搜索时间过长的问题。遗传算法原理：遗传算法是模拟达尔文生物进化论的自然选择和遗传学机理的生物进化过程的计算模

型，是一种通过模拟自然进化过程搜索最优解的方法。遗传算法能够自我迭代，让它本身系统内的东西进行优胜劣汰的自然选择，把好的保留下来，次一点的东西就排除掉。遗传算法的本质就是是优胜劣汰，选出最优秀的个体，一般用来寻找最优解。[3]

遗传算法是一种启发式算法，能够在相对于遍历搜索较短的时间内得到一个较优结果。由于启发式算法具有随机性，我们多次对该模型运行该算法，最终取平均。根据遗传算法得到的结果，在在温度为 387.8070℃，Co/SiO2 量为 177.33mg，HAP 量为 178.7992mg，Co 负载量为 1.3383wt%，乙醇浓度为 1.1442*ml/min* 时，C4 烯烃收率达到最高。

表 9 遗传算法运行结果

	温度	Co/SiO2 量	HAP 量	Co 负载量	乙醇浓度	C4 烯烃收率
1	387.0703	174.3541	193.6570	1.3095	1.1825	41.4497
2	385.2849	163.4369	198.6219	1.4314	1.1579	41.4497
3	386.4409	178.4056	167.7338	1.3239	1.0729	41.4497
4	394.4529	182.9676	165.5722	1.3245	1.1234	41.4497
5	385.7862	187.4858	168.4111	1.3021	1.1841	41.4497
平均值	387.8070	177.3300	178.7992	1.3383	1.1442	41.4497

5.3.5 将温度限制在 350℃以下考虑较优催化剂组合和温度

通过之前的数据图我们可以知道，350℃以前 C4 烯烃收率随温度变化的趋势与 350 摄氏度之后的有着较为显著的差异，从图上可以直观的看到：350℃到 400℃这段曲线的斜率明显陡于 350℃之前的斜率。所以在本问将温度限制在 350℃之下，我们需要重新清洗数据，找到该温度段的具体规律，也即将 350℃以后的数据全部抛弃，这样能够不受 350℃之后数据的影响，更高的拟合 350℃之前的数学规律。

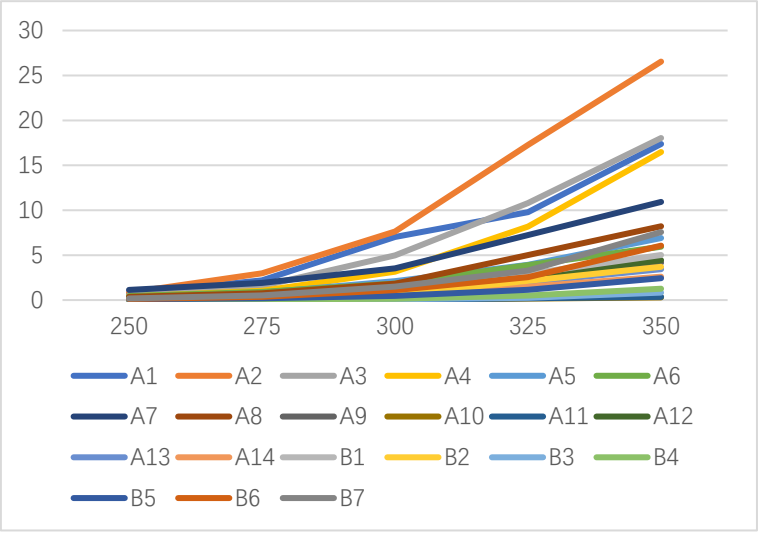


图 24　350℃各组 C4 烯烃收率随温度变化情况

此时研究的催化剂因素和温度的上下界如下表：

表 10 350℃以下各因素的搜索上下界

因素名	上界	下界	单位
Co 负载量	0.5	5	*wt%*
乙醇浓度	0.3	2.1	*ml/min*
温度	250	350	℃
Co/SiO2 的量	10	200	*mg*
HAP 的量	10	200	*mg*

仍然使用 AdaBoost、随机森林、XGBoost 三个模型来进行拟合，得到的结果再使用 VotingRegressor 进行集成，得到一个更合理的结果。

350℃之下的拟合 VotingRegressor 模型保存在支撑材料中。

同样在得到模型后，我们需要寻找全局最优值，由于遍历搜索的种种弊端，我们使用遗传算法进行求解，根据遗传算法的到的结果，我们可以认为将温度限制在 350℃之下时：温度为 341.4446℃，Co/SiO2 量为 176.3787*mg*，HAP 量为 178.2950mg，Co 负载量为 1.9306*wt%*，乙醇浓度为 1.6519*ml/min* 时，C4 烯烃收率达到最高。

表 11 遗传算法运行结果

	温度	Co/SiO2 量	HAP 量	Co 负载量	乙醇浓度	C4 烯烃收率
1	387.0703	174.3541	193.6570	1.3095	1.1825	41.4497
2	385.2849	163.4369	198.6219	1.4314	1.1579	41.4497
3	386.4409	178.4056	167.7338	1.3239	1.0729	41.4497
4	394.4529	182.9676	165.5722	1.3245	1.1234	41.4497
5	385.7862	187.4858	168.4111	1.3021	1.1841	41.4497
平均值	387.8070	177.3300	178.7992	1.3383	1.1442	41.4497

5.4 问题四的模型建立与求解

5.4.1 不同因素对 C4 烯烃收率影响的重要性的探究

我们增加实验来深入探究每种催化剂因素以及温度对 C4 烯烃收率的影响，但是由于实现增加次数的限制，我们必须有目的性的增加实验方案，否则探究将会走向低效甚至无效。新增实验的主要研究方法仍然是控制变量法，故我们需要固定一些自变量，而更改某些自变量来达到深入探究的目的。那么那种因素对 C4 烯烃收率的影响更大呢，我们使用了层次分析法(AHP)来进行探究。

层次分析法：根据问题的性质和要达到的目标，找到影响问题的不同影响因素，并按照因素间的相互关联影响以及隶属关系将因素按不同层次聚集组合，形成一个多层次的分析结构模型，从而最终使问题归结为最低层(供决策的方案、措施等)相对于最高层(总目标)的相对重要权值的确定或相对优劣次序的排定。[4]

首先我们对五个指标进行了重要性比较，即 Co 负载量、Co/SiO2 和 HAP

装料比、乙醇浓度、催化剂的量、温度这五个指标，根据以下比例标度表来描述重要的程度。

表 12 重要性描述比例标度表

因素 i 比因素 j	量化值
同等重要	1
稍微重要	3
较强重要	5
强烈重要	7
极端重要	9
两相邻判断的中间值	2，4，6，8

最终得到的判断矩阵如下：

表 13 五个因素的重要性判断性判断矩阵

	温度	催化剂的量	乙醇浓度	Co 负载量	装料比
温度	1.000	4.000	7.000	5.000	9.000
催化剂的量	0.250	1.000	5.000	1.000	7.000
乙醇浓度	0.143	0.200	1.000	0.250	2.000
Co 负载量	0.200	1.000	4.000	1.000	6.000
装料比	0.111	0.143	0.500	0.167	1.000

根据层次分析法的原理带入程序计算，得到的检验结果如下：

表 14 HAP 法检验结果

项	特征向量	权重值	最大特征值	CI 值
温度	2.652	53.033%		
催化剂的量	1.005	20.098%		
乙醇浓度	0.286	5.724%	5.225	0.056
Co 负载量	0.878	17.569%		
装料比	0.179	3.575%		

并对得到的结果进行一致性检验：

表 15 一致性检验结果

最大特征根	CI 值	RI 值	CR 值	一致性检验结果
5.225	0.056	1.120	0.050	通过

该判断结果通过一致性检验，可以认为我们对每种因素对 C4 烯烃收率的影响的重要性的比较判断合理。即得到了一种描述每个因素重要性的数学模型。

重要性排序如下：温度 > 催化剂的量 > 乙醇浓度 >Co 负载量 > 装料比。故我们在设计实验时需要按照这个重要性来优先验证排在前面的因素。如温度是

排在第一位的，所以我们在设计实验时，每一组实验都按照步长为 25℃，如实记录从 250℃到 650℃的所有数据。

5.4.2 对原始数据分布特征的研究

现有的实验数据大多提供了从 250℃至 400℃的实验结果，很少涉及到温度在 400℃以上的数据。我们在前一问进行数据的拟合时，因为数据量过少(只有一组数据中有)，我们直接去除了 400℃以上的数据。这样一来，在求解最优温度时，我们其实已经默认最优温度在 400℃以下了，但是实际上的最优温度有可能在 400℃以上。所以在增加的 5 次实验中，我们将收集更多的 400℃以上的实验数据，来更好地求解最优温度，正如上一节所讨论，每一组实验都按照步长为 25℃，如实记录从 250℃到 650℃的所有数据。

增加的部分数据一方面可以用来改善 VotingRegressor 的拟合，另一方面可以使得遗传算法的搜索区间变大，提高求解的合理性与准确性。

通过第三问最后的结果我们发现，无论有没有“350℃以下”这一限制条件，通过遗传算法解出的最优催化剂组合中 Co/SiO2 和 HAP 的装料比都非常接近 1:1（分别为 177.3300:178.7992 和 176.3787:178.2950）。究其原因，是因为我们在训练模型时使用的数据分布是非常不均衡的。对于 Co/SiO2 和 HAP 的装料比而言，绝大部分数据的装填比都是 1:1，是有很少一部分数据的装填比为其他数值。这样一来，我们的模型很少能学习到装填比为 1:1 之外的数据，这有可能会导致我们的模型泛化能力较差，并且模型收敛速度也会减慢。针对这个问题，我们打算增加一些装料比不为 1:1 的实验数据，以改善装料比的分布，改善模型的拟合效果。

相比之下，Co 负载量和乙醇浓度的数据分布偏差性要小一些，这从第三问求解出的结果也可以得到验证。在所有数据中，“Co 负载量为 1*wt%*”和“乙醇浓度为 1.68*ml/min*”是出现频率较高的两种数据，但是在第三问的结果中，Co 负载量和乙醇浓度的最优值与这两个指标的常见值还是有一定差距的，这也从侧面反映出 Co 负载量和乙醇浓度的数据分布偏差性相对较小。但是，从直观来看，“Co 负载量”和“乙醇浓度”的数据分布依然不是很均匀，可以适当增加一些实验数据来改善这种情况。

最后，由于装料方式对 C4 烯烃收率的影响不大，所以默认增加的实验均采用装料方式 I。

综合上述分析，我们增加的五组实验主要用来验证最优温度的求解和丰富装料比不为 1：1 的数据，以缓解装料比数据分布的不均衡性；在此基础上，适当的增加一些“Co 负载量”和“乙醇浓度”的非常见数值，改善其数据分布的不均衡性。

考虑增加下面的五组实验，具体每组实验的催化剂组合如下表所示。

表 16 新设计的 5 组实验的催化剂参数

实验编号	Co/SiO2 量（mg）	HAP 量（mg）	乙醇浓度（ml/min）	Co 负载量（1wt%）
C1	67	33	1.68	2

C2	100	50	1.68	1
C3	33	67	1.68	5
C4	67	33	0.9	1
C5	267	133	1.68	0.5

综合来看，这五组数据都尽可能地多提供温度超过 400 度的数据，以更好地求解能提高 C4 烯烃收率的最优温度。除此之外，这五组数据的 Co/SiO2 与 HAP 装料比为 2:1 或 1:2，可以较好地改善原本装料比数据分布不均衡的问题。我们可以把这些数据加入我们的机器学习模型中进行训练，观察数据拟合的效果是否会得到改善；并且可以将遗传算法求解的结果与加入数据之前的结果进行对比，观察关于装料比数据分布不均衡的问题是否得到改善。

5.4.3 新增实验的对比目的

表 17 各新增实验的目的

组别	催化剂参数	对比目的
C1	67mg 2wt%Co/SiO2-33mg HAP-乙醇浓度 1.68ml/min	与 A13 组实验（67mg 1wt%Co/SiO2-33mg HAP-乙醇浓度 1.68ml/min）进行对比，验证在 Co/SiO2 与 HAP 装料比为 2:1 时上一问中关于 Co 负载量的结论是否依然成立。
C2	100mg 1wt%Co/SiO2-50mg HAP-乙醇浓度 1.68ml/min	与 A13 组实验（67mg 1wt%Co/SiO2-33mg HAP-乙醇浓度 1.68ml/min）进行对比，验证在 Co/SiO2 与 HAP 装料比为 2:1 时上一问中关于催化剂量的结论是否依然成立。
C3	33mg 5wt%Co/SiO2-67mg HAP-乙醇浓度 1.68ml/min	与 A14 组实验（33mg 1wt%Co/SiO2-67mg HAP-乙醇浓度 1.68ml/min）进行对比，验证在 Co/SiO2 与 HAP 装料比为 1:2 时上一问中关于 Co 负载量的结论是否依然成立。
C4	67mg 1wt%Co/SiO2-33mg HAP-乙醇浓度	与 A13 组实验（67mg 1wt%Co/SiO2-33mg

	0.9ml/min	HAP- 乙醇浓度 1.68ml/min）进行对比，验证在 Co/SiO2 与 HAP 装料比为2:1时上一问中关于乙醇浓度的结论是否依然成立。
C5	267mg 0.5wt%Co/SiO2-133mg HAP-乙醇浓度 1.68ml/min	与A2组实验（200mg 2wt%Co/SiO2-200mg HAP- 乙醇浓度 1.68ml/min）进行对比，验证在 Co 负载量为 2wt%时上一问中关于 Co/SiO2 与 HAP 装料比的结论是否依然成立。

六、模型的评价与改进

6.1 评价

在第一问中，我们通过可视化的方式将因变量与自变量的关系直接画出，根据曲线走势(线性，二次，指数)利用 SPSS 软件对每一个曲线进行拟合，并采用拟合优度检验，判断所拟合的曲线较为合理，拟合效果比较显著。

在第二问中，我们采取控制变量法，对影响乙醇转化率和 C4 烯烃选择性的多种因素分别进行分析，得到了一系列结论。最后，通过 XGBoost 算法进行多元回归，对题中所给的部分数据进行拟合。XGBoost 算法是近些年提出的一种较为新颖的机器学习算法，在算法和工程层面对传统的 GBDT 算法进行了大量优化，性能优良。我们利用该算法对之前的结论进行检验，证明了结论的合理性。

在第三问中，我们采用了集成学习的方法，集成了 AdaBoost，随机森林，XGBoost 三种机器学习方法，建立 VotingRegressor 模型对题目所给的数据进行拟合。与单一的机器学习模型相比，我们的集成学习模型通用性更强，对数据的拟合效果也相当不错。

在第四问中，我们首先通过层次分析法对影响 C4 烯烃收率的各因素权重进行研究，并且针对集成学习时发现的数据分布不均衡的问题确定了新增的五组实验，这五组实验能够优先服务于权重较大的因素，并且能够较为有效地缓解数据分布不均衡的问题。

6.2 改进

机器学习效果的好坏在一定程度上取决于一些超参数的确定，由于时间有限，

我们的参数可能不是最优的，需要后续调整。

关于第四问中新增的五组实验，我们主要是针对我们发现的题目数据可能存在的缺陷进行补充，考虑得可能还不够全面。

七、参考文献

[1]吕绍沛. 乙醇偶合制备丁醇及 C_4 烯烃[D].大连理工大学,2018.

[2]Chen, T. and Guestrin, C., “XGBoost: A Scalable Tree Boosting System”, arXiv e-prints, 2016.

[3] 吴孟达，成礼智，吴翊,王丹，数学模型教程，北京：高等教育出版社,2013.

[4] 许树柏. 实用决策方法:层次分析法原理[M]. 天津大学出版社, 1988.

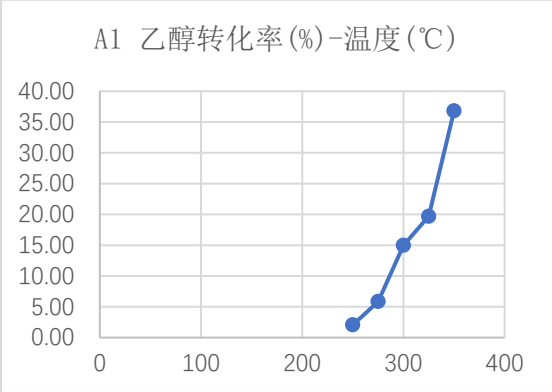
[5] 韩中庚.数学建模方法及应用[M].北京：高等教育出版社,2009.

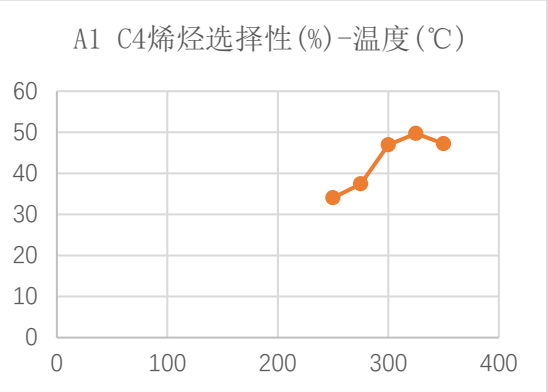
[6] 赵静，但琦.数学建模与数学实验[M].北京：高等教育出版社，2003（2006 重印）

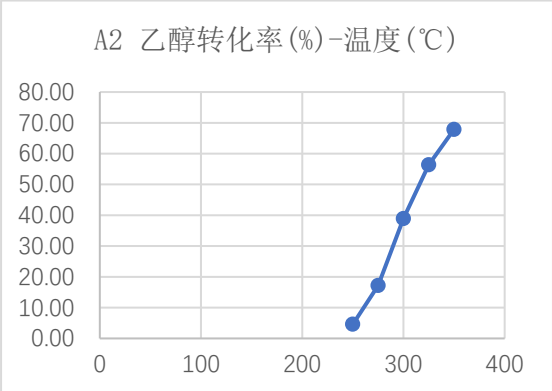
[7] 司守奎，孙兆亮，数学建模算法与应用，北京：国防工业出版社,2015.

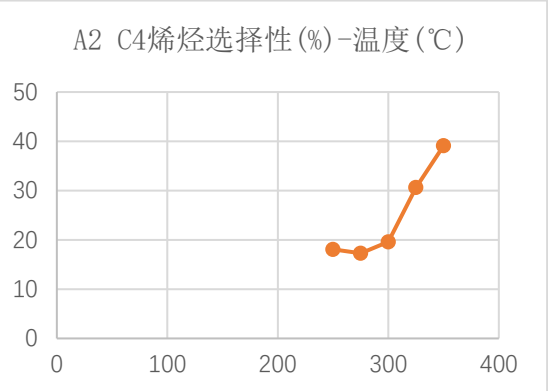
八、附录

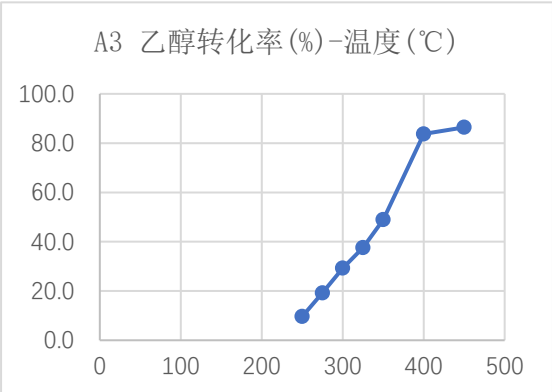
问题一中待拟合的各曲线的散点折线图

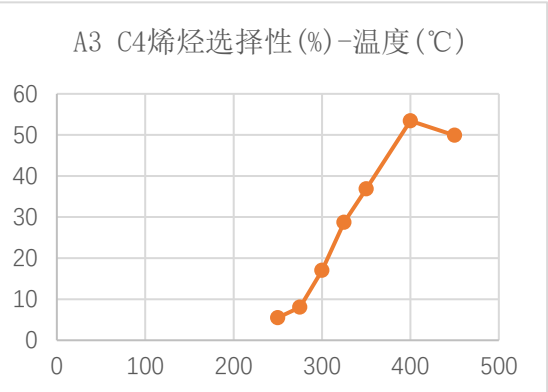


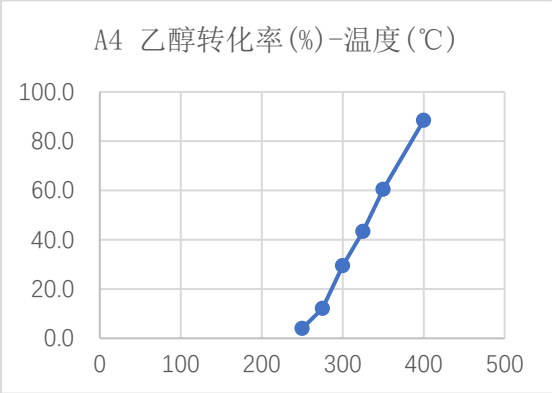


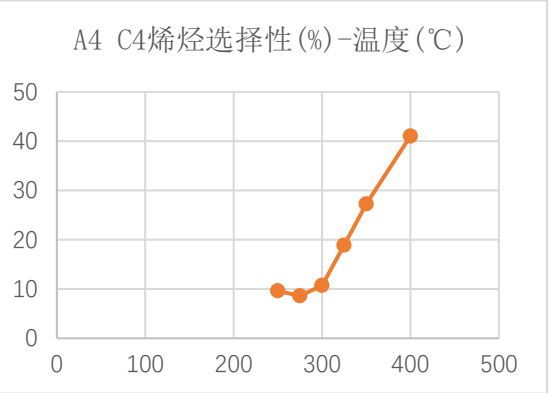


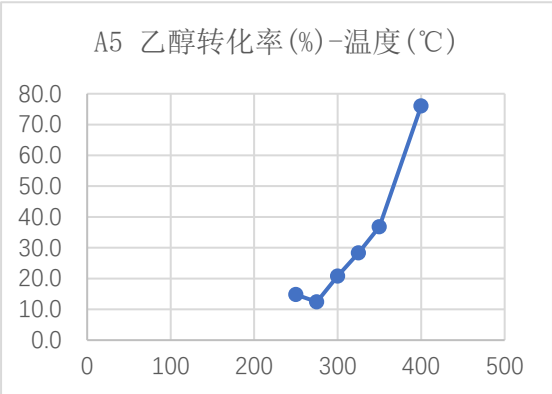


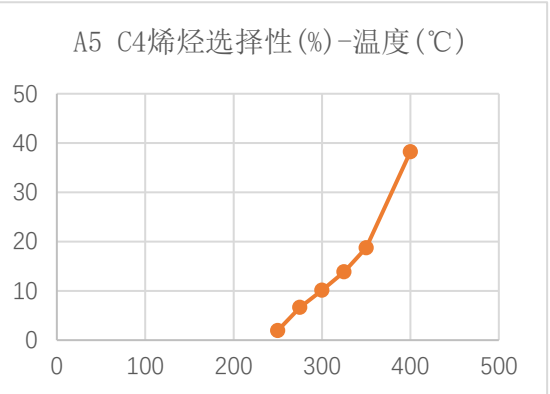


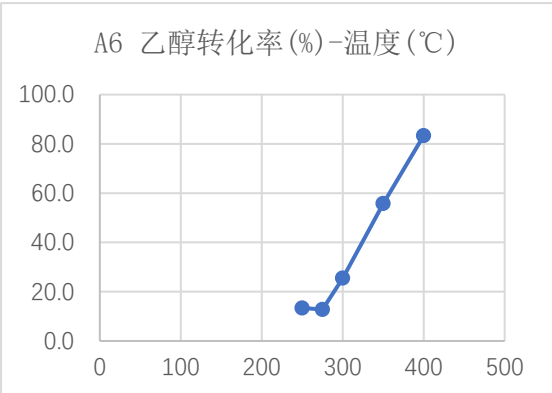


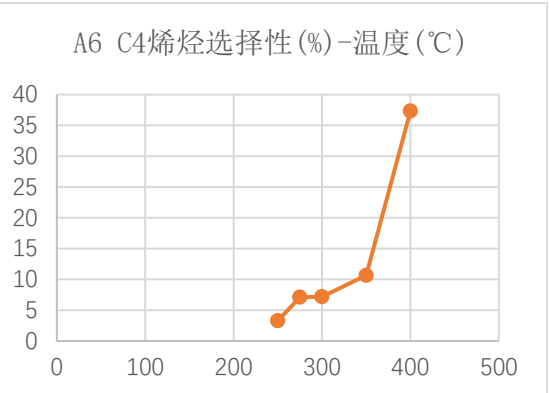

A4 乙醇转化率(%)-温度(℃)
100.0
80.0
60.0
40.0
20.0
0.0
0
100
200
300
400
500

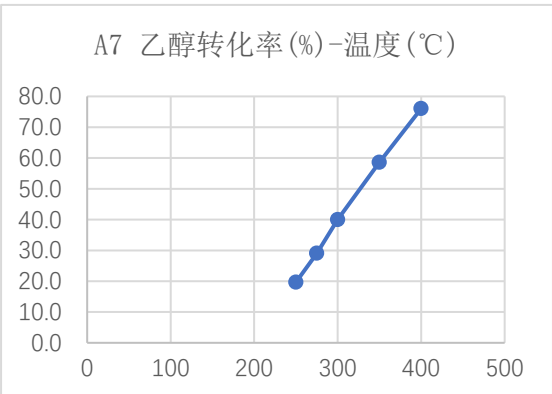

A4 C4烯烃选择性(%)-温度(℃)
50
40
30
20
10
0
0
100
200
300
400
500

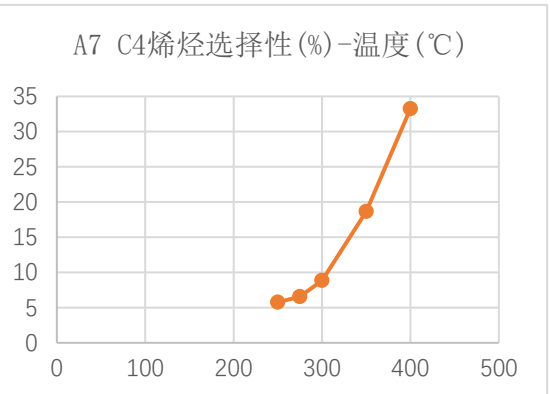

A5 乙醇转化率(%)-温度(℃)
80.0
70.0
60.0
50.0
40.0
30.0
20.0
10.0
0.0
0
100
200
300
400
500

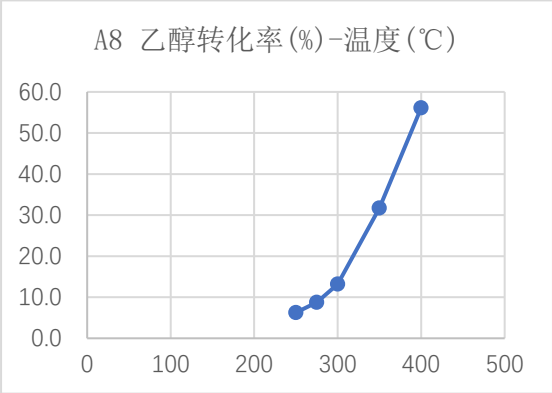

A5 C4烯烃选择性(%)-温度(℃)
50
40
30
20
10
0
0
100
200
300
400
500

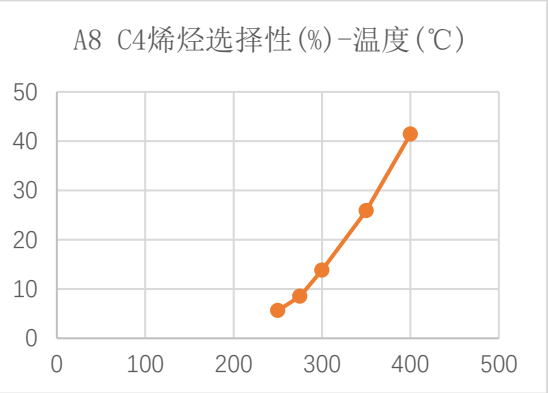

A6 乙醇转化率(%)-温度(℃)
100.0
80.0
60.0
40.0
20.0
0.0
0
100
200
300
400
500

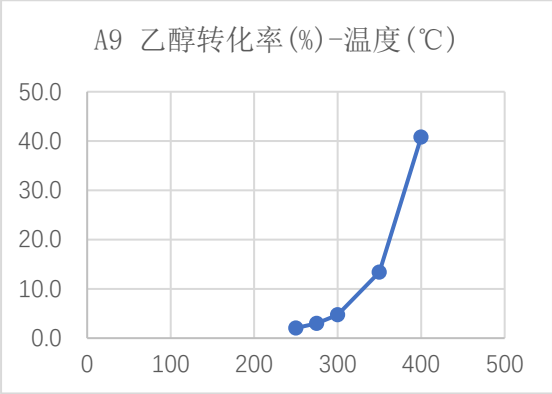

A6 C4烯烃选择性(%)-温度(℃)
40
35
30
25
20
15
10
5
0
0
100
200
300
400
500

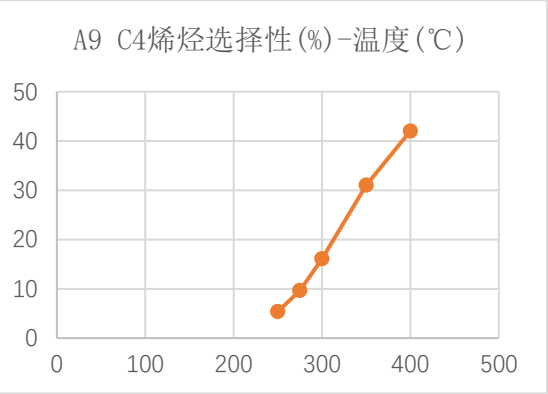

A7 乙醇转化率(%)-温度(℃)
80.0
70.0
60.0
50.0
40.0
30.0
20.0
10.0
0.0
0
100
200
300
400
500

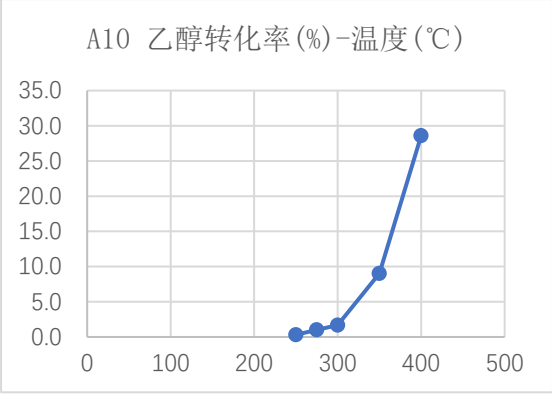

A7 C4烯烃选择性(%)-温度(℃)
35
30
25
20
15
10
5
0
0
100
200
300
400
500

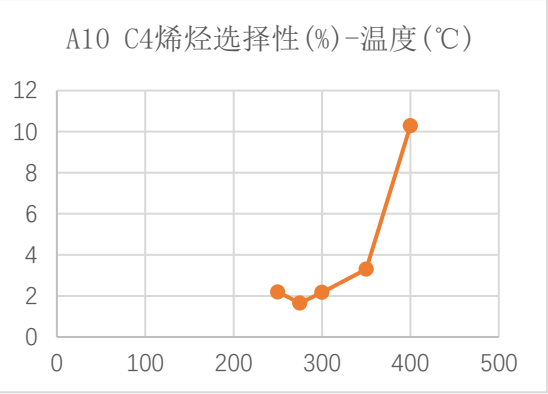

A8 乙醇转化率(%)-温度(℃)
60.0
50.0
40.0
30.0
20.0
10.0
0.0
0
100
200
300
400
500

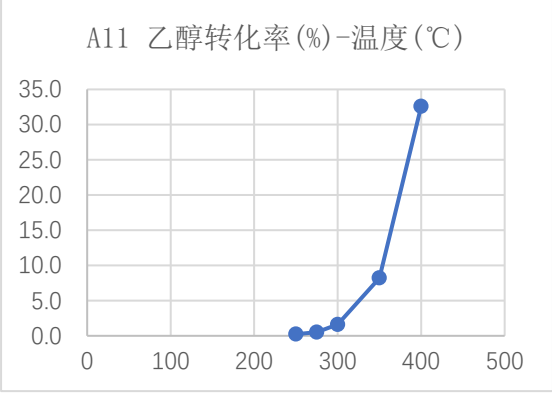

A8 C4烯烃选择性(%)-温度(℃)
50
40
30
20
10
0
0
100
200
300
400
500

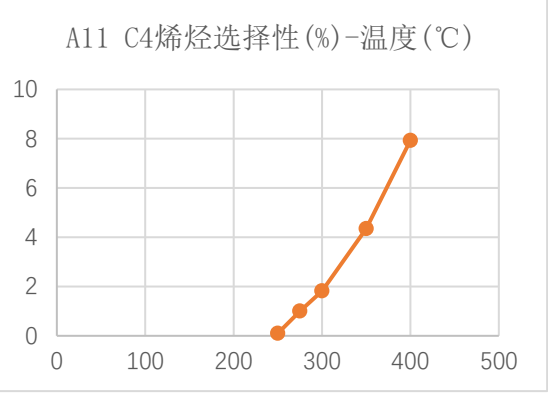

A9 乙醇转化率(%)-温度(℃)
50.0
40.0
30.0
20.0
10.0
0.0
0
100
200
300
400
500

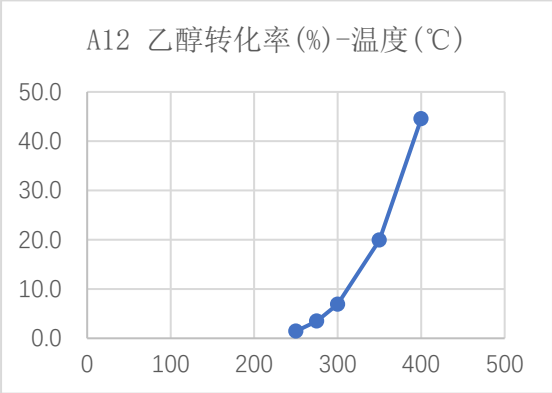

A9 C4烯烃选择性(%)-温度(℃)
50
40
30
20
10
0
0
100
200
300
400
500

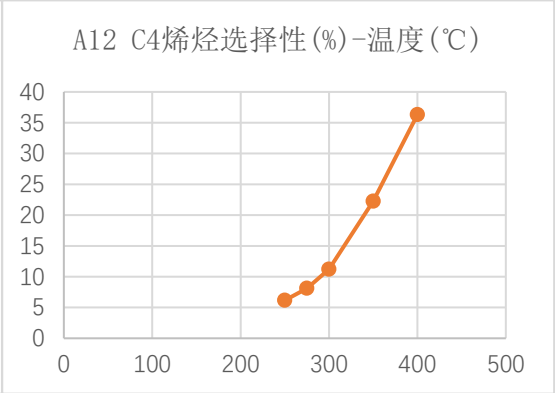

A10 乙醇转化率(%)-温度(℃)
35.0
30.0
25.0
20.0
15.0
10.0
5.0
0.0
0
100
200
300
400
500

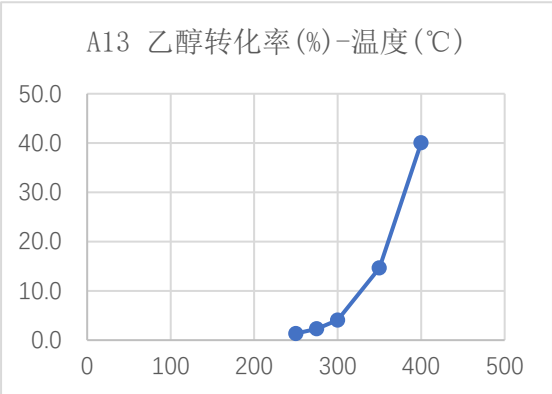

A10 C4烯烃选择性(%)-温度(℃)
12
10
8
6
4
2
0
0
100
200
300
400
500

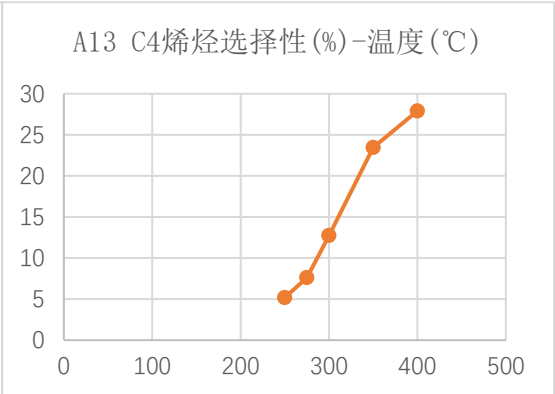

A11 乙醇转化率(%)-温度(℃)
35.0
30.0
25.0
20.0
15.0
10.0
5.0
0.0
0
100
200
300
400
500

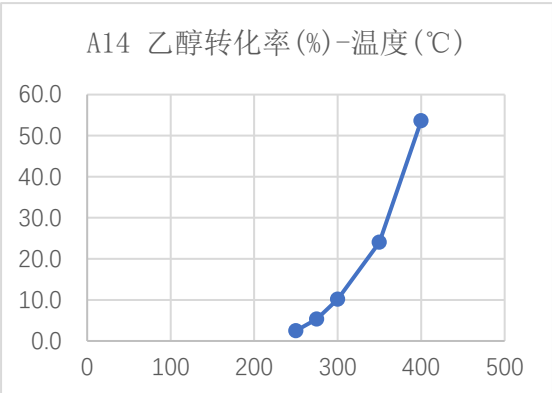

A11 C4烯烃选择性(%)-温度(℃)
10
8
6
4
2
0
0
100
200
300
400
500

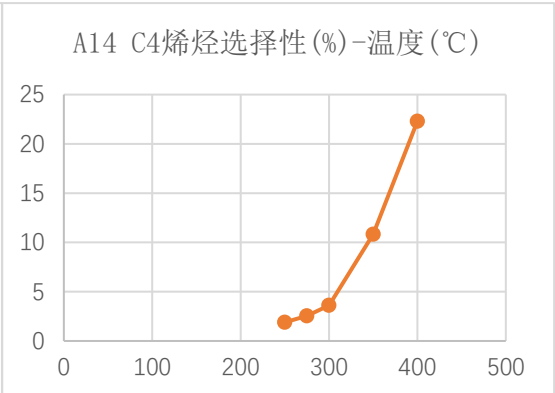

A12 乙醇转化率(%)-温度(℃)
50.0
40.0
30.0
20.0
10.0
0.0
0
100
200
300
400
500

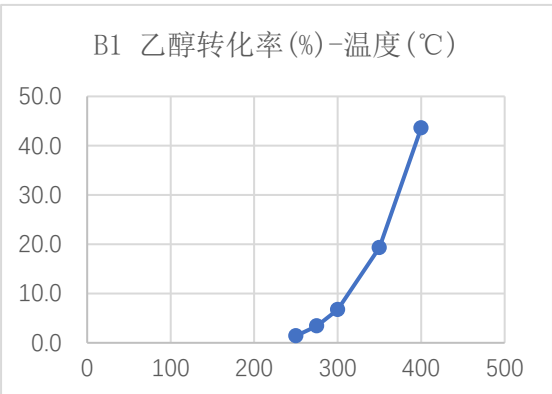

A12 C4烯烃选择性(%)-温度(℃)
40
35
30
25
20
15
10
5
0
0
100
200
300
400
500

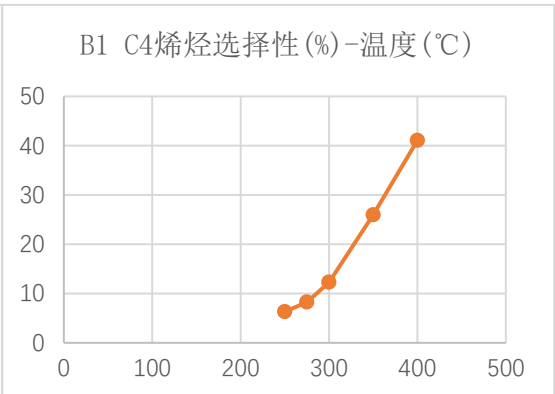

A13 乙醇转化率(%)-温度(℃)
50.0
40.0
30.0
20.0
10.0
0.0
0
100
200
300
400
500

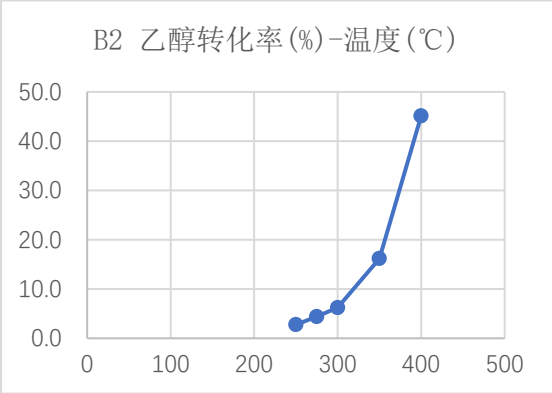

A13 C4烯烃选择性(%)-温度(℃)
30
25
20
15
10
5
0
0
100
200
300
400
500

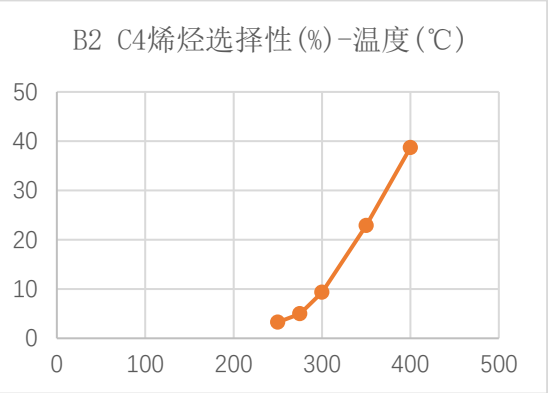

A14 乙醇转化率(%)-温度(℃)
60.0
50.0
40.0
30.0
20.0
10.0
0.0
0
100
200
300
400
500

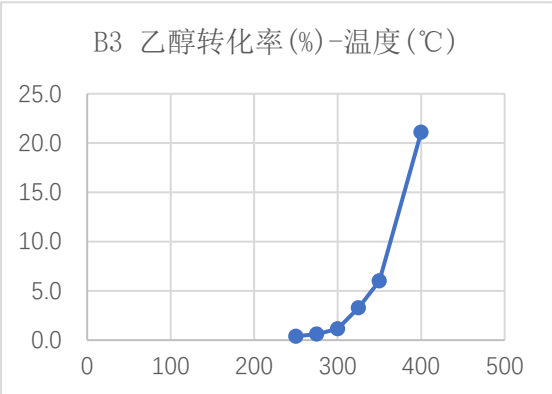

A14 C4烯烃选择性(%)-温度(℃)
25
20
15
10
5
0
0
100
200
300
400
500

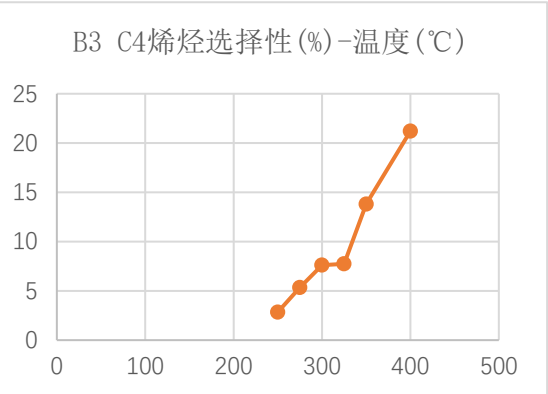

B1 乙醇转化率(%)-温度(℃)
50.0
40.0
30.0
20.0
10.0
0.0
0
100
200
300
400
500

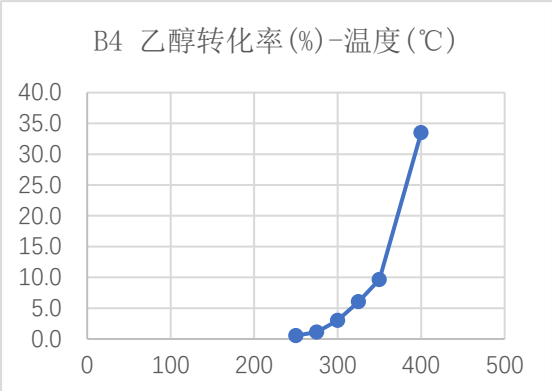

B1 C4烯烃选择性(%)-温度(℃)
50
40
30
20
10
0
0
100
200
300
400
500

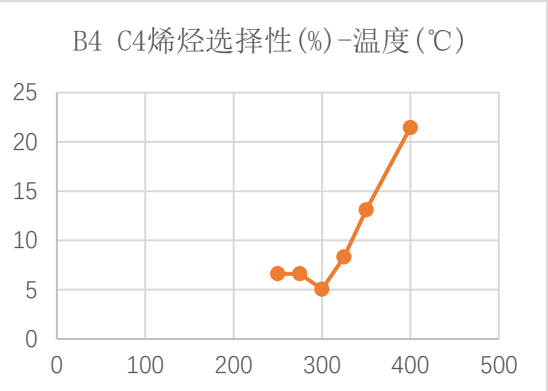

B2 乙醇转化率(%)-温度(℃)
50.0
40.0
30.0
20.0
10.0
0.0
0
100
200
300
400
500

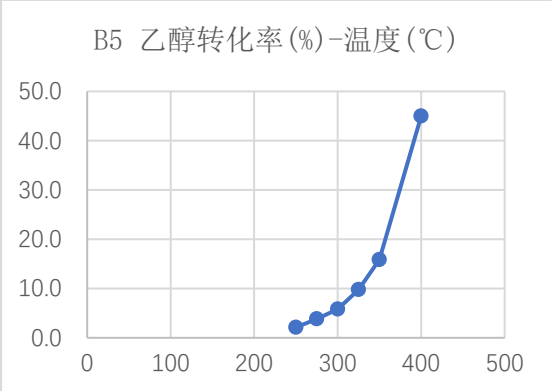

B2 C4烯烃选择性(%)-温度(℃)
50
40
30
20
10
0
0
100
200
300
400
500

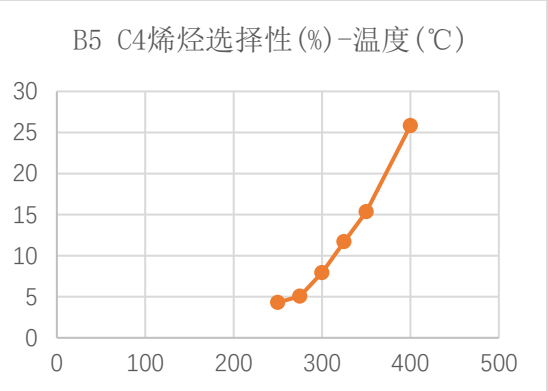

B3 乙醇转化率(%)-温度(℃)
25.0
20.0
15.0
10.0
5.0
0.0
0
100
200
300
400
500

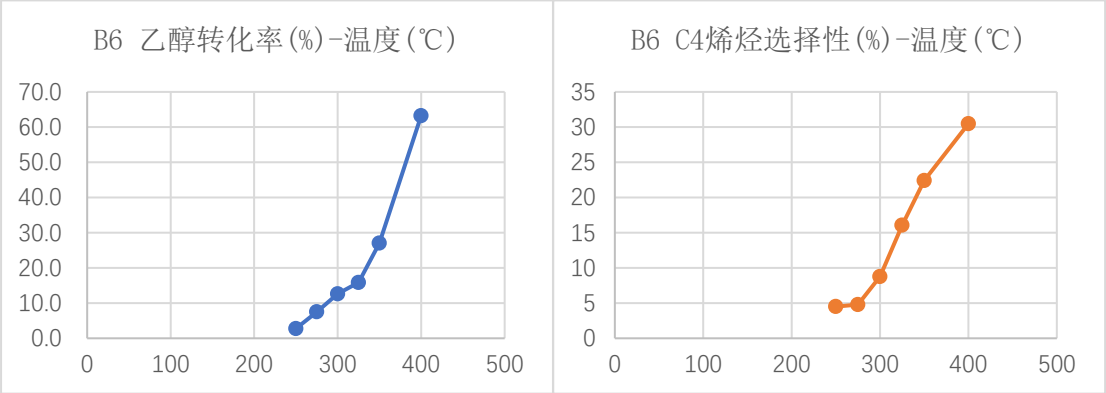

B3 C4烯烃选择性(%)-温度(℃)
25
20
15
10
5
0
0
100
200
300
400
500

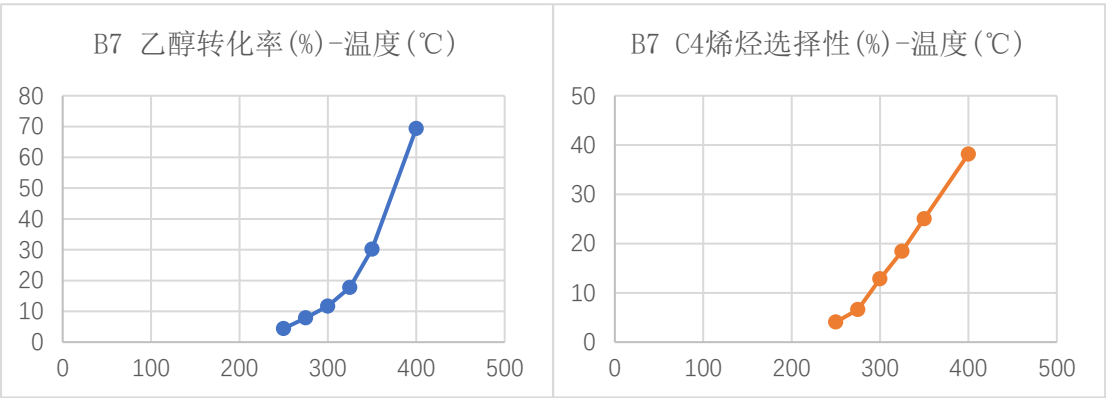

B4 乙醇转化率(%)-温度(℃)
40.0
35.0
30.0
25.0
20.0
15.0
10.0
5.0
0.0
0
100
200
300
400
500


B4 C4烯烃选择性(%)-温度(℃)
25
20
15
10
5
0
0
100
200
300
400
500


B5 乙醇转化率(%)-温度(℃)
50.0
40.0
30.0
20.0
10.0
0.0
0
100
200
300
400
500


B5 C4烯烃选择性(%)-温度(℃)
30
25
20
15
10
5
0
0
100
200
300
400
500





回归检验

乙醇转化率-温度：

组合编号	回归类型	R 方	F	自由度 1	自由度 2	显著性
A1	线性	.932	41.230	1	3	.008
A2	线性	.990	297.190	1	3	.000
A3	线性	.964	134.958	1	5	.000
A4	线性	.995	800.891	1	4	.000
A5	二次	.994	249.615	2	3	.000
A6	二次	.986	70.591	2	2	.014
A7	线性	.999	2468.722	1	3	.000
A8	二次	.999	1039.318	2	2	.001
A9	指数	.996	811.147	1	3	.000
A10	指数	.989	270.481	1	3	.000
A11	指数	.994	489.368	1	3	.000

A12	二次	.999	1089.424	2	2	.001
A13	指数	.998	1701.905	1	3	.000
A14	二次	.997	351.384	2	2	.003
B1	二次	.999	889.039	2	2	.001
B2	指数	.997	1077.689	1	3	.000
B3	指数	.991	429.404	1	4	.000
B4	指数	.988	326.097	1	4	.000
B5	指数	.999	3932.252	1	4	.000
B6	二次	.990	152.328	2	3	.001
B7	指数	.997	1543.744	1	4	.000

C4 选择性-温度：

组合编号	回归类型	R 方	F	自由度 1	自由度 2	显著性
A1	线性	.787	11.079	1	3	.045
A2	线性	.836	15.286	1	3	.030
A3	线性	.913	52.349	1	5	.001
A4	线性	.917	44.368	1	4	.003
A5	线性	.940	62.768	1	4	.001
A6	线性	.784	10.886	1	3	.046
A7	线性	.937	44.907	1	3	.007
A8	线性	.983	175.899	1	3	.001
A9	线性	.995	574.575	1	3	.000
A10	二次	.978	45.444	2	2	.022
A11	线性	.978	134.718	1	3	.001
A12	线性	.967	86.856	1	3	.003
A13	线性	.977	126.165	1	3	.002

A14	线性	.920	34.486	1	3	.010
B1	线性	.972	103.253	1	3	.002
B2	线性	.970	96.326	1	3	.002
B3	线性	.943	65.903	1	4	.001
B4	二次	.974	55.500	2	3	.004
B5	线性	.956	86.374	1	4	.001
B6	线性	.965	108.856	1	4	.000
B7	线性	.989	353.933	1	4	.000

SPSS

线性回归

```
TSET NEWVAR=NONE.
CURVEFIT
  /VARIABLES=Y WITH X
  /CONSTANT
  /MODEL=LINEAR
  /PLOT FIT.
```

二次回归

```
TSET NEWVAR=NONE.
CURVEFIT
  /VARIABLES=Y WITH X
  /CONSTANT
  /MODEL=QUADRATIC
  /PLOT FIT.
```

指数回归

```
TSET NEWVAR=NONE.
CURVEFIT
  /VARIABLES=Y WITH X
  /CONSTANT
  /MODEL=EXPONENTIAL
  /PLOT FIT.
```

幂回归

```
TSET NEWVAR=NONE.
CURVEFIT
  /VARIABLES=lambd WITH Y350
  /CONSTANT
  /MODEL=POWER
  /PLOT FIT.
```

方差分析

```
ONEWAY 乙醇转化率 C4 烯烃选择性 BY 装料方式
  /STATISTICS DESCRIPTIVES HOMOGENEITY
  /MISSING ANALYSIS
  /POSTHOC=LSD ALPHA(0.05).
```

Python

AHP 层次分析法

```
import numpy as np

A = np.array([[1, 4, 7, 5, 9],
              [1/4, 1, 5, 1, 7],
              [1/7, 1/5, 1, 1/4, 2],
              [1/5, 1, 4, 1, 6],
              [1/9, 1/7, 1/2, 1/6, 1]])

lens = len(A)
RI = [0, 0, 0.52, 0.89, 1.12, 1.26, 1.36, 1.41, 1.46, 1.49, 1.52, 1.54, 1.56, 1.58, 1.59]
R = np.linalg.matrix_rank(A)  # 求A的秩
V, D = np.linalg.eig(A)  # 求A的特征值和特征向量
V = list(V)
_MAX = np.max(V)  # 计算最大特征值
index = V.index(_MAX)
C = D[:, index]
CI = (_MAX - lens) / (lens - 1)  # 计算一致性检验指标CI
```

```
CR = CI / RI[lens]
if CR < 0.10:
    print('已通过一致性检验，各向量权重向量 Q 为：')
    Q = C / np.sum(C)  # 特征向量标准化
    print(Q)  # 输出权重向量
else:
print("未通过一致性检验")
```

XGBRegressor 模型

train.py

```
import numpy as np
import pandas as pd
from xgboost import XGBRegressor
from sklearn.model_selection import train_test_split
import joblib

df_raw = pd.read_excel(r'C:\Users\hp\Desktop\处理后的数据.xlsx')

X1 = df_raw.loc[df_raw['温度'] <= 400, '温度'].values.reshape(-1, 1)
X2 = df_raw.loc[df_raw['温度'] <= 400, ['Co/SiO2 量', 'HAP 量', 'Co 负载量', '乙醇浓度']].values
X = np.hstack([X1, X2])
y1 = df_raw.loc[df_raw['温度'] <= 400, '乙醇转化率(%)'].values
y2 = df_raw.loc[df_raw['温度'] <= 400, 'C4 烯烃选择性(%)'].values

X_train1, X_test1, y_train1, y_test1 = train_test_split(X, y1, random_state=50, train_size=0.87)
X_train2, X_test2, y_train2, y_test2 = train_test_split(X, y2, random_state=60, train_size=0.87)

print('研究目标：乙醇转化率')
xgb_reg = XGBRegressor(n_estimators=150)
xgb_reg.fit(X_train1, y_train1)
print("XGB 训练得分：")
print(xgb_reg.score(X_test1, y_test1))
joblib.dump(xgb_reg, 'C2H6O.model')
```

```
24.
25. print('研究目标：C4 烯烃选择性')
26. xgb_reg = XGBRegressor(n_estimators=150)
27. xgb_reg.fit(X_train2, y_train2)
28. print("XGB 训练得分：")
29. print(xgb_reg.score(X_test2, y_test2))
30. joblib.dump(xgb_reg, 'C4.model')
```

XGB 训练分数见下表。

乙醇转化率	C4 烯烃收率
0.954136390842374	0.939745171224007

GA 遗传算法

train400.py

```
1. import numpy as np
2. import pandas as pd
3. from sklearn.ensemble import RandomForestRegressor
4. from sklearn.ensemble import AdaBoostRegressor
5. from xgboost import XGBRegressor
6. from sklearn.ensemble import VotingRegressor
7. from sklearn.model_selection import train_test_split
8. import joblib
9.
10. df_raw = pd.read_excel(r'C:\Users\hp\Desktop\处理后的数
    据.xlsx')
11.
12. X1 = df_raw.loc[df_raw['温度'] <= 400, '温度
    '].values.reshape(-1, 1)
13. X2 = df_raw.loc[df_raw['温度'] <= 400, ['Co/SiO2 量
    ', 'HAP 量', 'Co 负载量', '乙醇浓度']].values
14. X = np.hstack([X1, X2])
15. y = df_raw.loc[df_raw['温度'] <= 400, 'C4 烯烃收率
    '].values
16.
17. X_train, X_test, y_train, y_test = train_test_split(X, y
    , random_state=50, train_size=0.87)
18.
19. rf_reg = RandomForestRegressor(n_estimators=150)
```

```
rf_reg.fit(X_train, y_train)
print("RF 训练得分：")
print(rf_reg.score(X_test, y_test))

xgb_reg = XGBRegressor(n_estimators=150)
xgb_reg.fit(X_train, y_train)
print("XGB 训练得分：")
print(xgb_reg.score(X_test, y_test))

ab_reg = AdaBoostRegressor(n_estimators=150)
ab_reg.fit(X_train, y_train)
print("AB 训练得分：")
print(ab_reg.score(X_test, y_test))

vo_reg = VotingRegressor(estimators=[('rf', rf_reg), ('xgb', xgb_reg), ('ab', ab_reg)])
vo_reg.fit(X_train, y_train)
print("Vo 训练得分：")
print(vo_reg.score(X_test, y_test))

# vo 是一个 VotingRegressor 模型
joblib.dump(vo_reg, 'vo400.model')
```

VotingRegressor 训练分数（400 度）见下表。

RF 训练得分	XGB 训练得分	AB 训练得分	Vo 训练得分
0.935274075677155	0.921863717197563	0.941335604311528	0.946041295029638

将模型保存在 vo400.model 文件中。

train350.py

```
import numpy as np
import pandas as pd
from sklearn.ensemble import RandomForestRegressor
from sklearn.ensemble import AdaBoostRegressor
from xgboost import XGBRegressor
from sklearn.ensemble import VotingRegressor
from sklearn.model_selection import train_test_split
import joblib

df_raw = pd.read_excel(r'C:\Users\hp\Desktop\处理后的数据.xlsx')
```

```
11.
12. X1 = df_raw.loc[df_raw['温度'] <= 350, '温度
    '].values.reshape(-1, 1)
13. X2 = df_raw.loc[df_raw['温度'] <= 350, ['Co/SiO2 量
    ', 'HAP 量', 'Co 负载量', '乙醇浓度']].values
14. X = np.hstack([X1, X2])
15. y = df_raw.loc[df_raw['温度'] <= 350, 'C4 烯烃收率
    '].values
16.
17. X_train, X_test, y_train, y_test = train_test_split(X, y
    , random_state=60, train_size=0.87)
18.
19. rf_reg = RandomForestRegressor(n_estimators=150)
20. rf_reg.fit(X_train, y_train)
21. print("RF 训练得分：")
22. print(rf_reg.score(X_test, y_test))
23.
24. xgb_reg = XGBRegressor(n_estimators=150)
25. xgb_reg.fit(X_train, y_train)
26. print("XGB 训练得分：")
27. print(xgb_reg.score(X_test, y_test))
28.
29. ab_reg = AdaBoostRegressor(n_estimators=150)
30. ab_reg.fit(X_train, y_train)
31. print("AB 训练得分：")
32. print(ab_reg.score(X_test, y_test))
33.
34. vo_reg = VotingRegressor(estimators=[('rf', rf_reg), ('x
    gb', xgb_reg), ('ab', ab_reg)])
35. vo_reg.fit(X_train, y_train)
36. print("Vo 训练得分：")
37. print(vo_reg.score(X_test, y_test))
38.
39. # vo 是一个 VotingRegressor 模型
40. joblib.dump(vo_reg, 'vo350.model')
```

VotingRegressor 训练分数（350 度）见下表。

RF 训练得分	XGB 训练得分	AB 训练得分	Vo 训练得分
0.983842452161144	0.921994277760717	0.874686757968277	0. 958546341965074

将模型保存在 vo350.model 文件中。

GA.py

```
1. import matplotlib.pyplot as plt
2. import numpy as np
3. import pandas as pd
4. import joblib
5.
6. def schaffer(p):
7.     '''
8.     This function has plenty of local minimum, with stron
   g shocks
9.     global minimum at (0,0) with value 0
10.     '''
11.     x1, x2, x3, x4, x5 = p
12.     return -
   vo_reg.predict(np.array([x1, x2, x3, x4, x5]).reshape((-
   1, 5))).item()
13.
14.
15. if __name__ == '__main__':
16.     vo_reg = joblib.load('vo350.model')
17.
18.     # 遗传算法
19.
20.     from sko.GA import GA
21.
22.     ga = GA(func=schaffer, n_dim=5, size_pop=50, max_ite
   r=30, prob_mut=0.01, lb=[250, 10, 10, 0.5, 0.3],
23.             ub=[350, 200, 200, 5, 2.1], precision=1e-7)
24.     best_x, best_y = ga.run()
25.     print('best_x:', best_x, '\n', 'best_y:', best_y)
26.
27.     # Plot the result
28.
29.     plt.rcParams['font.sans-serif'] = ['SimHei']
30.     plt.rcParams['axes.unicode_minus'] = False
31.
32.     Y_history = pd.DataFrame(ga.all_history_Y)
33.     Y_history.min(axis=1).cummin().plot(kind='line')
34.     plt.xlabel('迭代次数')
35.     plt.ylabel('-C4 烯烃收率(%)')
36.     plt.show()
```

算法结果具体如下（即 result.xlsx）
无温度条件限制：

序号	best_x					best_y
	温度	Co/SiO2 量	HAP 量	Co 负载量	乙醇浓度	C4 烯烃收率
1	387.0703	174.3541	193.6570	1.3095	1.1825	41.4497
2	385.2849	163.4369	198.6219	1.4314	1.1579	41.4497
3	386.4409	178.4056	167.7338	1.3239	1.0729	41.4497
4	394.4529	182.9676	165.5722	1.3245	1.1234	41.4497
5	385.7862	187.4858	168.4111	1.3021	1.1841	41.4497
均值	387.8070	177.3300	178.7992	1.3383	1.1442	41.4497

有温度条件限制：

序号	best_x					best_y
	温度	Co/SiO2 量	HAP 量	Co 负载量	乙醇浓度	C4 烯烃收率
1	341.2118	191.5445	162.2371	1.9712	1.4319	21.7147
2	338.9285	174.3598	179.4177	2.1757	1.5678	21.7147
3	340.4001	198.7576	166.0325	2.1308	1.6716	21.7147
4	342.5169	154.2124	197.9212	1.6640	1.7444	21.7147
5	344.1658	163.0193	185.8666	1.7115	1.8438	21.7147
均值	341.4446	176.3787	178.2950	1.9306	1.6519	21.7147

支撑材料内容组成

文件夹	文件名	主要功能/用途
代码	线性回归.sps	利用已有数据点进行线性回归（SPSS 源代码）
	二次回归.sps	利用已有数据点进行二次回归（SPSS 源代码）
	指数回归.sps	利用已有数据点进行指数回归（SPSS 源代码）
	幂回归.sps	利用已有数据点进行幂回归（SPSS 源代码）
	方差分析.sps	针对不同的装料方式进行方差分析对比（SPSS 源代码）
	AHP.py	对温度及催化剂组合中的几个变量进行层次分析，计算各因素的权重大小（python 源代码）
	GA.py	利用遗传算法对保存的机器学习预测函数进行最优值求解（python 源代码）
	train.py	针对第二问对温度及催化剂组合各变量关于乙醇转

		化率和 C4 烯烃选择性的机器学习模型进行训练（python 源代码）
	train350.py	第三问 350 度以下对温度及催化剂组合各变量关于 C4 烯烃收率的机器学习模型进行训练（python 源码）
	train400.py	第三问无温度条件限制时对温度及催化剂组合各变量关于 C4 烯烃收率的机器学习模型进行训练（python 源码）
模型	C2H6O.model	第二问拟合出的温度及催化剂组合各变量关于乙醇转化率的机器学习模型（MODEL 文件）
	C4.model	第二问拟合出的温度及催化剂组合各变量关于 C4 烯烃选择性的机器学习模型（MODEL 文件）
	vo350.model	第三问 350 度以下拟合出的温度及催化剂组合各变量关于 C4 烯烃收率的机器学习模型（MODEL 文件）
	vo400.model	第三问无温度条件限制时拟合出的温度及催化剂组合各变量关于 C4 烯烃收率的机器学习模型（MODEL 文件）
数据	data.sav	利用 SPSS 软件进行回归拟合、方差分析等操作时使用的数据（SPSS 数据集）
	result.xlsx	保存着遗传算法多次测试的结果（EXCEL 文件）
	处理后的数据.xlsx	处理后的实验数据（EXCEL 文件）

说明：

（1）在用 SPSS 进行统计分析时，首先需要导入 data.sav 数据集。在进行线性、指数、二次回归分析时，需要将待分析的数据拷贝至数据集中 X 和 Y 变量对应的位置，替代原有的数据进行分析。进行幂回归和方差分析时不必导入新的数据。

（2）在运行 GA.py 前，首先需要通过下列语句安装一些依赖库：

pip install matplotlib numpy pandas scikit-opt joblib

（3）在运行 train.py，train350.py，train400.py 前，需要将“处理后的数据.xlsx”拷贝到桌面上，并根据实际文件路径修改代码里的下列语句

df_raw = pd.read_excel(r'C:\Users\hp\Desktop\处理后的数据.xlsx')

将引号里的部分修改为文件的实际位置。

另外需要通过下列语句安装一些依赖库：

pip install numpy pandas xgboost sklearn joblib

（4）在运行上述所有 python 代码前，要把 MODEL 文件拷贝至与 python 源代码的相同目录下。