|  |  |  |
| --- | --- | --- |
| Problem Chosen C | 2023 MCM/ICM Summary Sheet | Team Control Number 1111111 |

**Abstract**

Contents

[Introduction 3](#_Toc127555886)

[1.1 Background 3](#_Toc127555887)

[1.2 Problem Statement and Analysis 3](#_Toc127555888)

[2. General Assumptions and Justification 3](#_Toc127555889)

[3. Variable Description 3](#_Toc127555890)

[4. Task 1 3](#_Toc127555891)

[4.1 Conclusion 3](#_Toc127555892)

[4.2 Problem Analysis 3](#_Toc127555893)

[4.3 Model Building 3](#_Toc127555894)

[4.4 Discussion 3](#_Toc127555895)

[5. Task 2 3](#_Toc127555896)

[5.1 Conclusion 3](#_Toc127555897)

[5.2 Problem Analysis 3](#_Toc127555898)

[5.3 Model Building 3](#_Toc127555899)

[5.4 Discussion 3](#_Toc127555900)

[6. Task 3 3](#_Toc127555901)

[6.1 Conclusion 4](#_Toc127555902)

[6.2 Problem Analysis 4](#_Toc127555903)

[6.3 Model Building 4](#_Toc127555904)

[6.4 Discussion 4](#_Toc127555905)

[7. Task 4 4](#_Toc127555906)

[7.1 Conclusion 4](#_Toc127555907)

[7.2 Problem Analysis 4](#_Toc127555908)

[7.3 Model Building 4](#_Toc127555909)

[7.4 Discussion 4](#_Toc127555910)

[8. Memo 4](#_Toc127555911)

[9. Reference 4](#_Toc127555912)

[10. Appendix 4](#_Toc127555913)

# Introduction

## Background

## Problem Statement and Analysis

# General Assumptions and Justification

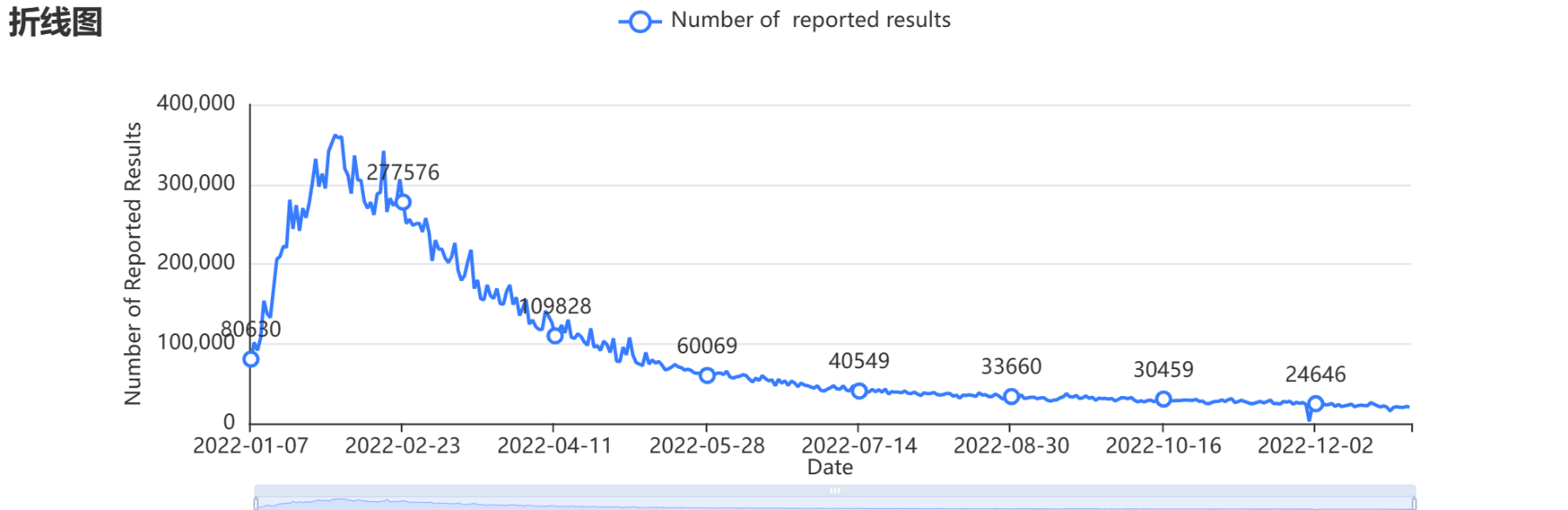
# Variable Description

# Task 1

## Conclusion

Number of reported results on 1, March, 2023 is **20626-20623**, the predicted result is  **20626.8285329377** by lightgbm Regression.

Number of reported results on 1, March, 2023 is **20539-20540**, the predicted result is **20539.842** by XGboost Regression.



**Figure 01** Date-Number of Reported Results

**Figure 02** Prediction of Number of Reported Results by Lightgbm

**Figure 03** Prediction of Number of Reported Results by XGBoost

## Problem Analysis

The reason why we choose Xgboost and Lightgbm will be stated in 4.4 Discussion Part.

## Model Building

**Lightgbm Model**

|  |  |
| --- | --- |
| 参数名 | 参数值 |
| 训练用时 | 0.212s |
| 数据切分 | 0.9 |
| 数据洗牌 | 是 |
| 交叉验证 | 10 |
| 基学习器 | gbdt |
| 基学习器数量 | 100 |
| 学习率 | 0.1 |
| L1正则项 | 0 |
| L2正则项 | 1 |
| 样本征采样率 | 1 |
| 树特征采样率 | 1 |
| 节点分裂阈值 | 0 |
| 叶子节点中样本的最小权重 | 0 |
| 树的最大深度 | 10 |
| 叶子节点最小样本数 | 10 |



**Figure 04**测试数据预测图

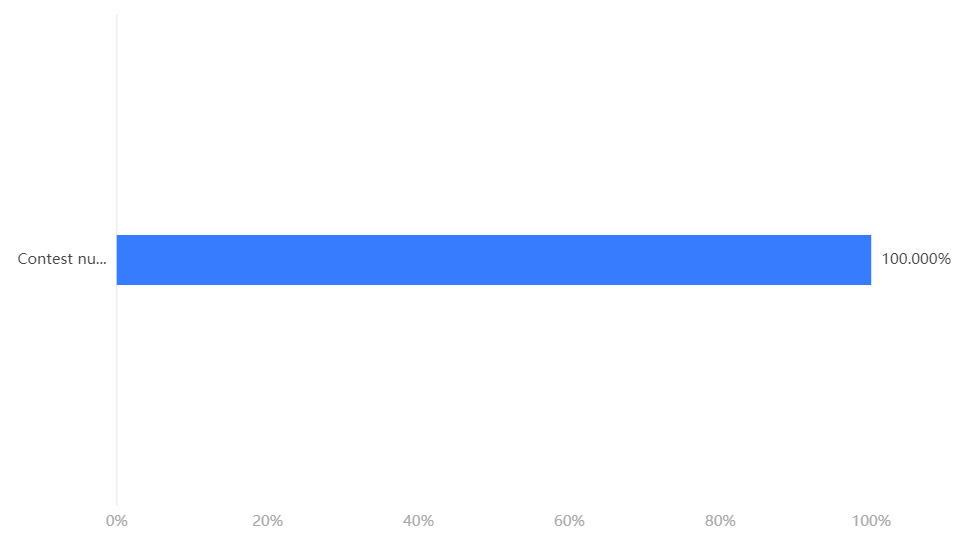


Figure 05特征重要性

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | MSE | RMSE | MAE | MAPE | R² |
| 训练集 | 182560768.337 | 13511.505 | 5645.762 | 5.615 | 0.977 |
| 交叉验证集 | 333263871.477 | 16104.582 | 7945.388 | 7.265 | 0.959 |
| 测试集 | 195690134.304 | 13988.929 | 8259.073 | 6.659 | 0.974 |

|  |  |
| --- | --- |
| 评价指标 | 评价结果 |
| MSE | 148104275.46996272 |
| RMSE | 12169.810001391259 |
| MAE | 6600.744028575751 |
| R² | 0.9813651722125035 |
| MAPE | 8.712185811616443 |

**XGboost Model**

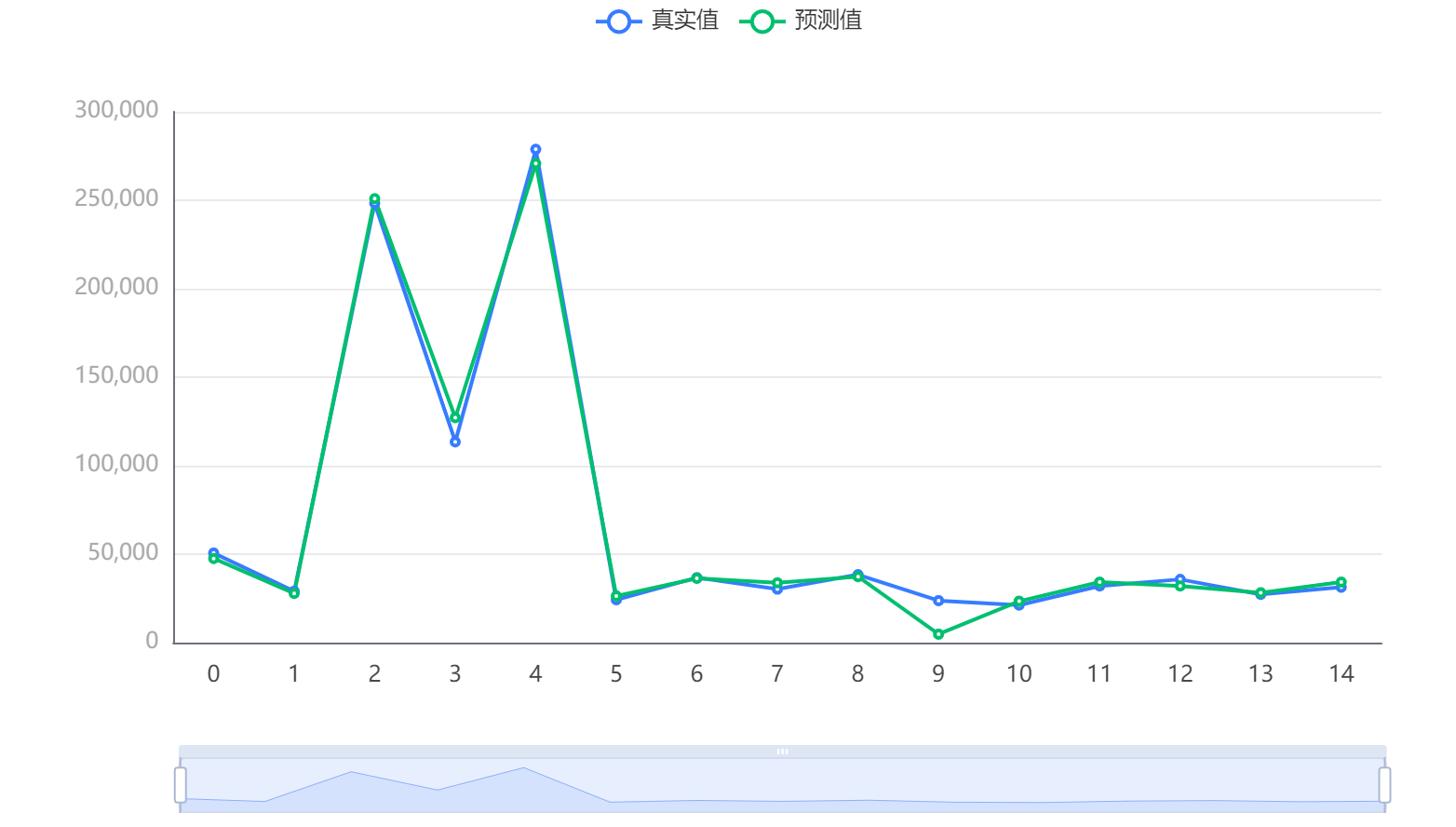


Figure 06测试数据预测图

**模型参数**

|  |  |
| --- | --- |
| 参数名 | 参数值 |
| 训练用时 | 1.162s |
| 数据切分 | 0.9 |
| 数据洗牌 | 是 |
| 交叉验证 | 10 |
| 基学习器 | gbtree |
| 基学习器数量 | 100 |
| 学习率 | 0.1 |
| L1正则项 | 0 |
| L2正则项 | 1 |
| 样本征采样率 | 1 |
| 树特征采样率 | 1 |
| 节点特征采样率 | 1 |
| 叶子节点中样本的最小权重 | 0 |
| 树的最大深度 | 10 |

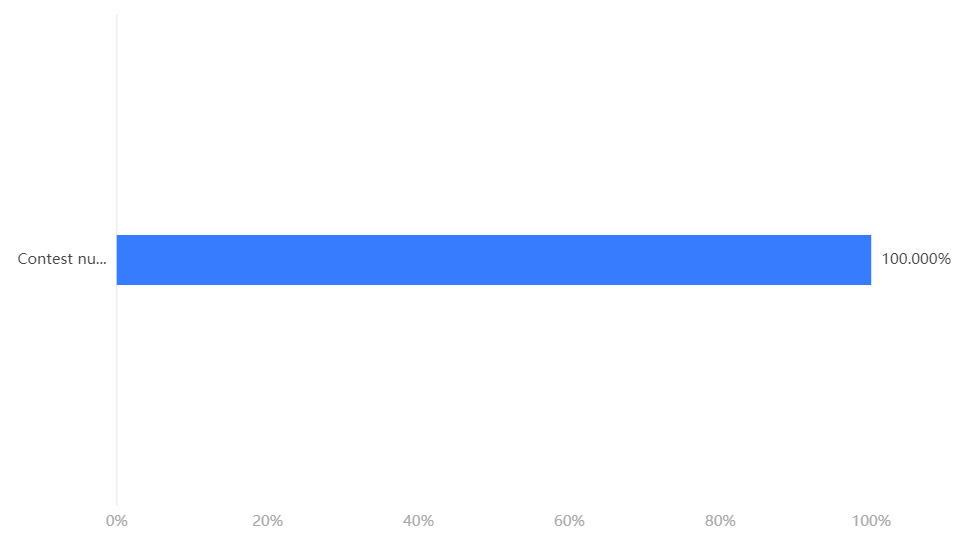


Figure 07特征重要性

**模型评估结果**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | MSE | RMSE | MAE | MAPE | R² |
| 训练集 | 1118762.369 | 1057.716 | 758.608 | 1.377 | 1 |
| 交叉验证集 | 160396274.429 | 12180.939 | 6996.049 | 6.845 | 0.98 |
| 测试集 | 82254400.256 | 9069.421 | 5101.003 | 17.305 | 0.984 |

## Discussion

Our initial choice to use polynomial regression has many advantages, such as a wide range of functions that can accommodate it, polynomials that are basically suitable for a wide range of curvatures, and polynomials that provide the best approximation of the relationship between dependent and independent variables. However, the disadvantages of using polynomial regression are that they are too sensitive to outliers, and the existence of one or two outliers in the data can seriously affect the results of nonlinear analysis. In addition, unfortunately, fewer model validation tools are available to detect outliers in nonlinear regression than in linear regression. In this problem, it is difficult to find patterns to predict. We first use the grey prediction model and found that the error was very large. Then we considered using polynomial fitting. We import the data and used MATLAB to fit the ninth order polynomial function but find that the predicted value is obviously incorrect. In addition, it is also difficult to study the ninth order polynomial function.

In addition, exponential regression has a large error in the early stage, it has a good fit in the later stage, but the predicted result is not reasonable. We try to use machine learning including xgboost and LightGBM to predict, the predicted result is reasonable.

Here we give our work on Polynomial Prediction and Exponent Prediction:

**Number of reported results on 1, March, 2023**

|  |  |
| --- | --- |
| Polynomial (Probably Wrong) | Result |
| Middle Bound |  |
| Lower Bound |  |
| Upper Bound |  |

|  |  |
| --- | --- |
| Exponent (Probably Right) | Result |
| Prediction | 7013 |

**Polynomial Prediction:**

图表, 折线图

描述已自动生成

图形用户界面, 应用程序, Word

描述已自动生成

**Figure 08** Polynomial Model Fitting Prediction Graph

**Exponent Prediction:**

**Figure 10** Exponent Model Fitting Prediction Graph

**Polynomial Model:**

Our team use polynomial model fitting prediction graph to fit the curve, and findings as follows:

***Middle Bound:***

***Lower Bound:***

***Upper Bound:***

|  |  |
| --- | --- |
| Goodness of fit: |  |
| *SSE:* |  |
| *R-square:* | 0.996 |
| *Adjusted R-square:* | 0.9959 |
| *RMSE:* | 5740 |

**Exponent Model:**

# Task 2

## Conclusion

## Problem Analysis

## Model Building

## Discussion

# Task 3

## Conclusion

## Problem Analysis

## Model Building

## Discussion

# Task 4

## Conclusion

## Problem Analysis

## Model Building

## Discussion

# Memo

# Reference

# Appendix