

ALY6110: Data Management & Big Data

College Of Professional Studies

**Final Project – Basic Analysis and Dashboard**

**COVID-19 Deaths**

By: Group 2 (Divya Bomaraboina,Sindhu Reddy Naini,Zhenyu Zhang)

Instructor: Hema Seshadri

November 25, 2024

### **Introduction**

The dataset under analysis, sourced from the CDC, is a comprehensive record of COVID-19 deaths from January 2020 to September 2023. The primary objective of this analysis is to leverage historical death data to predict future COVID-19 deaths and offer actionable recommendations to governments and hospitals. This dataset holds a wide range of features, including temporal data ("Year" and "Month"), geographic data ("State"), demographic attributes ("Sex" and "Age Group"), and critical health statistics such as "COVID-19 Deaths," "Total Deaths," "Pneumonia Deaths," and other comorbidity data. This extensive scope enables multifaceted insights into the dynamics of the pandemic.

Our goal is twofold: first, to identify patterns and correlations within the dataset that influence COVID-19 mortality rates, and second, to predict trends and offer guidance for resource allocation and policy interventions. The analysis emphasizes understanding the interplay between diseases like pneumonia and influenza with COVID-19 to prioritize healthcare strategies effectively.

The dataset’s richness in time, demographic, and disease-specific dimensions makes it a strong candidate for predictive modeling and analysis. By segmenting data into subsets by year, month, and total, we can derive deeper insights into temporal and regional trends. This report serves as a draft overview of the methodology, initial findings, and proposed visualizations. It highlights significant factors like high mortality rates in specific age groups and states and correlates these to policy and resource implications. With predictive modeling, this study aims to assist stakeholders in optimizing healthcare strategies to address ongoing and future pandemic challenges.

**Exploratory Data Analysis**

The initial step in the analysis was to understand the structure and composition of the dataset. The dataset contains 137,700 rows and 16 columns, making it a large and comprehensive resource for studying COVID-19-related mortality. By examining the data types of each column, it was determined that the dataset includes both categorical and numerical data, such as demographic attributes ("State," "Sex," and "Age Group") and critical health indicators ("COVID-19 Deaths," "Pneumonia Deaths," "Influenza Deaths," etc.).

**Data Cleaning and Transformation**

To prepare the data for meaningful analysis, several preprocessing steps were applied:

1. **Null Values:** Rows containing null values were removed to ensure data integrity and reliability.
2. **Duplicate Records:** Duplicate entries were dropped to avoid overrepresentation of data.
3. **Irrelevant Columns:** Columns deemed unnecessary, such as those with uniform values or unrelated metadata, were removed.
4. **Segmentation:** The data was divided into three subsets based on the "Group" column:
   * **By Total:** Aggregated statistics without temporal granularity.
   * **By Year:** Yearly statistics for a broader temporal overview.
   * **By Month:** Monthly data for more detailed temporal trends.

文本

描述已自动生成

Figure 1.1

图片包含 图示

描述已自动生成

Figure 1.2

#### **Correlation Analysis and Heatmap**

To explore relationships between variables, a correlation heatmap was generated. This visual representation highlighted several key patterns:

* Strong correlations were observed between "COVID-19 Deaths" and other disease-related metrics, particularly "Pneumonia Deaths" and "Pneumonia and COVID-19 Deaths." This indicates a high comorbidity effect between COVID-19 and pneumonia, emphasizing the need for targeted healthcare interventions for respiratory diseases.
* Weaker correlations were noted with demographic features such as "State," "Sex," and "Age Group," suggesting that while these factors are relevant, they may not be as directly predictive of mortality as disease-related metrics.

To handle categorical variables like "State" and "Sex," one-hot encoding was applied, converting these categories into numerical values while preserving their inherent distinctions. This transformation enabled a more nuanced analysis and compatibility with machine learning models.

图表

描述已自动生成

Figure 2.1

The subset "df1" (grouped by total) was analyzed to assess state-level differences. California, Texas, Florida, Pennsylvania, and Ohio emerged as the top five states with the highest total deaths. This can be attributed to their large populations and high urban density, which amplify the risk of virus transmission. These insights visualized using a choropleth map, offering a clear spatial representation of mortality trends.

In "df2" (grouped by year), demographic patterns explored. While males showed slightly higher mortality rates than females, this disparity is due to population differences rather than any direct biological predisposition. Age analysis revealed that older populations, especially those aged 75 and above, faced the highest mortality risks, consistent with weakened immune responses in elderly individuals.

地图

描述已自动生成

Figure 2.2

#### **Temporal Trends**

The "df3" (grouped by month) provided a granular look at how COVID-19 deaths fluctuated over time. Peaks were observed in January 2021 and January 2022, which can be linked to:

1. **Seasonality:** Cold weather weakens immune systems, making individuals more susceptible to respiratory infections.
2. **Social Gatherings:** Increased interactions during holiday seasons likely contributed to transmission spikes.

The decline in deaths from mid-2022 onward aligns with widespread vaccination campaigns and improved treatment protocols. However, despite the decrease in COVID-19 deaths, diseases like pneumonia continued to record significant mortality rates, underscoring the need for sustained healthcare focus on respiratory conditions.

**Key Observations**

* **Comorbidities:** The high correlation between "COVID-19 Deaths" and "Pneumonia Deaths" highlights the critical role of managing comorbid conditions in reducing mortality.
* **Population Density:** High-population states require more robust healthcare infrastructure and intervention plans.
* **Elderly Care:** Older adults remain a vulnerable demographic requiring prioritized healthcare services.
* **Seasonal Preparedness:** Policies need to address seasonal spikes, particularly during winter months and holiday periods.

These exploratory findings not only validate the dataset’s utility but also lay a strong foundation for predictive modeling and actionable insights. Visualizations such as heatmaps, choropleth maps, and temporal plots enriched the analysis, providing clarity and direction for further research.

图表, 条形图

描述已自动生成

Figure 2.3

图表

描述已自动生成

Figure 2.4

The figure above shows how age related with Deaths, and we can clearly see that the old people is, the more death there is. The reason could be that the old people will have worse immune system, so it is easier for them to death.

Now, let’s go to df3 group by month to see how time affects COVID-19 Deaths. The figure below shows that the first and second climax is in 2021-01 and 2022-01. This might have two reasons, first is because the cold weather reduces people’s immune system which makes death number increase. Next is because there are many festivals around January, so there are many people meet tighter which might also increase death number. Also, we can se that is the death number is extremely decrease after middle of 2022 and get close to 0 in 2023. But for disease like Pneumonia, although its death number is also decrease as COVID-19 death decrease, its death number is still large.

图表, 折线图

描述已自动生成

Figure 2.6

图表, 折线图

描述已自动生成

Figure 2.7

**Insight**

Now, with all the information above, we can see that first, the trend of death of COVID-19 death is keep reducing now, and it is almost go to 0. Thus, for government hospitals, they can spend less money and energy on the COVID-19 problem, and for the public, they don’t need to be afraid of COVID-19 anymore. Instead, all government, hospital, and public need to take care of pneumonia which have large death number than COVID-19 now.

Next, consider the relation between COVID-19 and other features, we can see that the flourishing state or state with higher population have more death. Thus, these states need to pay more attention to COVID-19 than other states. Next, as we show that sex have no relation with COVID-19, but age has large effect. The old people always need more attention and more help not only on COVID-19, but also on pneumonia as they lack immune system to against them. Also, time has a large effect as well. When the weather is cold like January in winter, people need to take care of themselves to keep warm to improve immune power and keep away from crowded places because the risk of transmission of disease will increase. Last but not least, we can clearly see that pneumonia has an extremely high linear relationship with COVID-19, we think this might be because they are all respiratory diseases. The low relation between influenza and COVID-19 proof it as well, thus, pneumonia is the disease that government and hospital need to take care of as it will increase and decrease when COVID-19 increase or decrease. And its deaths number is even a little bit larger than COVID-19 when COVID-19 goes almost to 0.

**Modeling**

To predict the future death, I do four different model to see which model is best for this dataset. The first model I chose is a simple linear dataset, its performance is that Mean Squared Error is 82.42049762225037 and R-squared is 0.9891601928591927. Then, let’s look at its Cross-Validated MSE is 99.77869593375976. And its residual plot is shown in Figure 3.1.1. Thus, we can see that although the performance of model looks good. However, its residual plot shows that the model doesn’t perform well on test datasets. By considering such high performance and worse residual plot, I think the problem is overfitting which means that the model fit train dataset too well that it loses the ability of generalization. To solve the overfitting problem, I tried L1 and L2 method

Thus, I choose Ridge and Lasso model which is also linear model to see which model is a better improvement for linear model. Based on the results of Cross-Validated MSE (Ridge) is 1.510271994860769, and Cross-Validated MSE (Lasso) is 1.5358253069445076. I chose Ridge model to improve linear model, and the performance is that Mean Squared Error (Ridge) is 1.520949127860919 while R-squared (Ridge) is 0.9997999672934207. And RMSE is 1.233267662699756. Thus, we can see that the performance improved a lot. However, the residual plot for updated model in figure 3.1.2 shows that the problem of underfitting has been improved but not solved which means that the linear model can’t fit this dataset perfectly.

图表, 散点图

描述已自动生成

Figure 3.1.1

图表, 散点图

描述已自动生成

Figure 3.1.2

The next model I choose is random forest model, this is a tree model different from linear model before, it can show some non-linear relation between different columns. Thus, it should fit model better and give a better residual plot. The performance of RF model is Mean Squared Error is 8.791593286627121 and R-squared is 0.9988437442330884. RMSE is 2.965062105020251 and the Cross-validated MSE is 13.042910890148272. We can see that though performance is better than original linear model but worse than Ridge model. And the residual plot in Figure 3.2.1 shows that it fits better with datasets than all linear models. Next, I do hyperparameter tuning for RF model, and the result I get is that Best Cross-Validated MSE is 123.18656446929113, Optimized Mean Squared Error is 100.84344694932958, and Optimized R-squared is 0.9867372371208559. The performance is much worse than the initial RF model, and the residual plot in Figure 3.2.2 for improved RF model is also has a few more outliers. Thus, based on the performance of RF model, we get conclusion that Tree model does better than Linear model on residual plot but has less performance than linear model. Based on this idea, I chose the GBM model which is tree model usually with better performance than RF and it is also the best traditional model in my experience.

图表, 散点图

描述已自动生成

Figure 3.2.1

图表, 散点图

描述已自动生成

Figure 3.2.2

The last traditional I choose is GBM, its performance is Mean Squared Error is 26.598299155986606, R-squared is 0.9965018358121808, and Cross-validated MSE is 28.57541184512715. We can see that its performance is worse than initial RF model, and it is residual plot also doesn’t have a lot of improvement as shown in Figure 3.3.1. But after I do hyperparameter tuning, it gets a big improvement. Its Best Cross-Validated MSE is 6.6972207975294555, Optimized Mean Squared Error is 6.964972264887826, and Optimized R-squared is 0.999083978400149. RMSE is 2.587898915632034 and its residual plot is shown in figure 3.3.2. We can see that both its performance and residual plot is better than RF model, and while its performance is a little bit worse than Ridge model, its residual plot is much better than it. As a result, my conclusion is that GBM is the best model for this dataset.

图表, 散点图

描述已自动生成

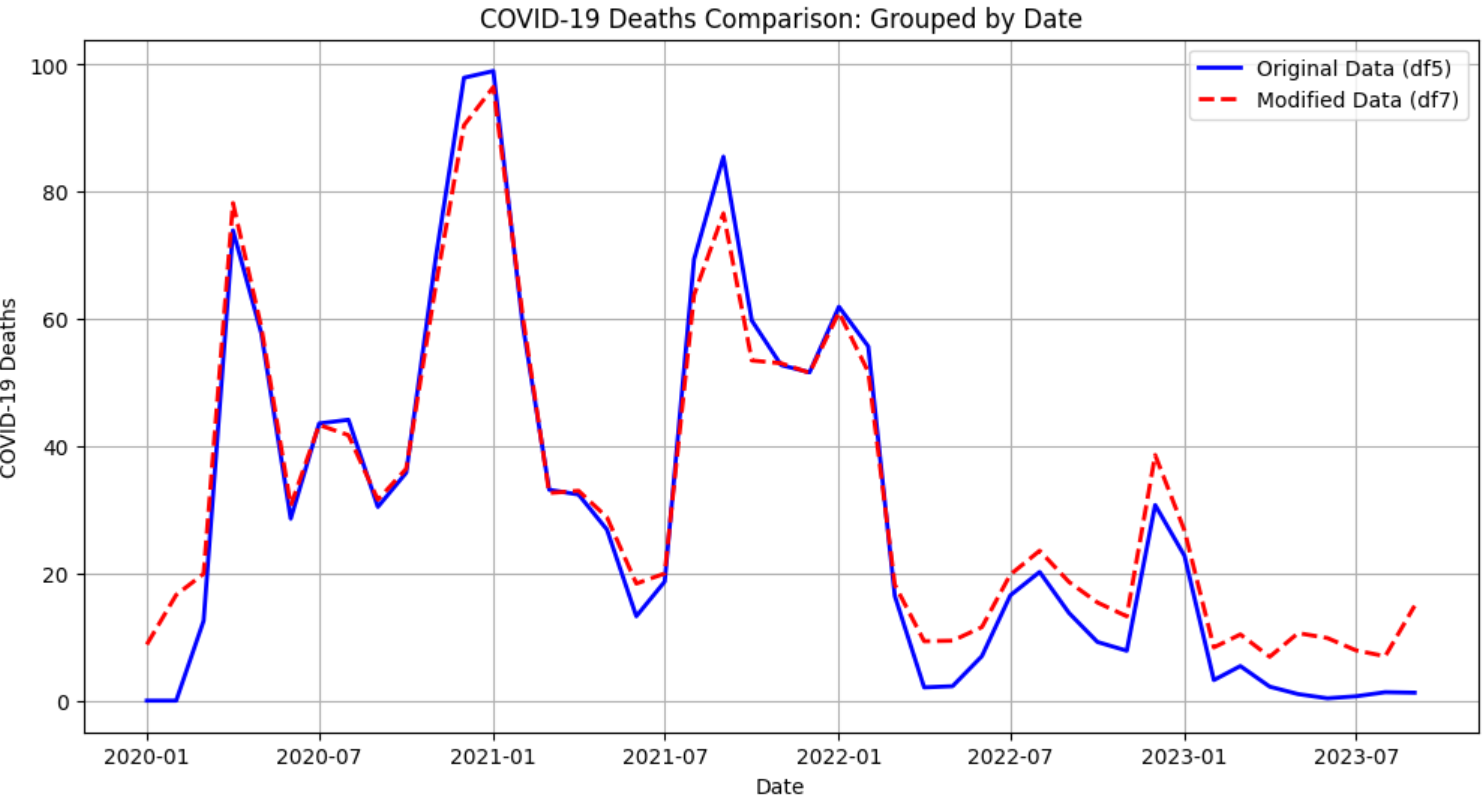
Figure 3.3.1

图表, 散点图

描述已自动生成

Figure 3.3.2

Besides these traditional models above, I also try a deep learning model which is LSTM model. I have done 30 epochs, and its performance is Mean Squared Error is 66.14676826523599 and R-squared is 0.9927328175627637 while its Average Cross-Validation MSE is 100.85001027361311. It is not good at all for a deep learning model, so I do a hyperparameter tuning for the model. And the performance is Mean Squared Error is 12.771035268984182 and R-squared is 0.9985969164383067, the RMSE is 3.5736585271936914 while Average Cross-Validation MSE is 29.877840274301462. Thus, we can see that deep learning can’t provide a better performance than GBM. I think the first reason is that the epochs is not enough, and the method to improve performance is not enough, or maybe this dataset is simple enough to use traditional model like GBM to show the relationship between feature, deep learning is more focus on non-linear relationship. And the figure 3.4.1 is the plot of comparing predict results of deep learning with real results in dataset. We can see that when death number is high, the model can fit very well, however, when death number go to 0 like recent trend, the model can’t show the future trend correctly.



**Conclusion**

In this dataset, we found that location with more people, old people, and respiratory diseases are all the reasons that cause large number of COVID-19 deaths. Although the trend of COVID-19 is almost go to 0 nowadays, pneumonia and the comorbidities disease with COVID-19 are not go to 0. Thus, we suggest that the government, hospital, and public can spend less energy on COVID-19, but more time on pneumonia or the comorbidities like pneumonia and COVID-19.

**References**

Cai, J., Xu, K., Zhu, Y., Hu, F., & Li, L. (2020). Prediction and analysis of net ecosystem carbon exchange based on gradient boosting regression and random forest. Applied energy, 262, 114566. <https://doi.org/10.1016/j.apenergy.2020.114566>

Camacho, J. (2014). *Visualizing big data with compressed score plots: approach and research challenges*. Chemometrics and Intelligent Laboratory Systems, 135, 110-125. <https://doi.org/10.1016/j.chemolab.2014.04.011>

Ioannidis, J.P.A. (2021). *Over- and under-estimation of COVID-19 deaths*. Eur J Epidemiol 36, 581–588. <https://doi.org/10.1007/s10654-021-00787-9>

Melkumova, L. E., & Shatskikh, S. Y. (2017). Comparing Ridge and LASSO estimators for data analysis. Procedia engineering, 201, 746-755. <https://doi.org/10.1016/j.proeng.2017.09.615>

Mortensen, E. M., Coley, C. M., Singer, D. E., Marrie, T. J., Obrosky, D. S., Kapoor, W. N., & Fine, M. J. (2002). *Causes of death for patients with community-acquired pneumonia: results from the Pneumonia Patient Outcomes Research Team cohort study*. Archives of internal medicine, 162(9), 1059-1064. <https://doi/10.1001/archinte.162.9.1059>

Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M., & Chica-Rivas, M. J. O. G. R. (2015). Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. Ore Geology Reviews, 71, 804-818. <https://doi.org/10.1016/j.oregeorev.2015.01.001>

Sherstinsky, A. (2020). Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. Physica D: Nonlinear Phenomena, 404, 132306. <https://doi.org/10.1016/j.physd.2019.132306>

U.S. Department of Health & Human Services. (2023, September 29). *U.S. Department of Health & Human Services - Provisional Covid-19 deaths by sex and age*. Catalog. <https://catalog.data.gov/dataset/provisional-covid-19-death-counts-by-sex-age-and-state>

**Appendix**

地图

描述已自动生成

地图

描述已自动生成

地图

描述已自动生成

地图

描述已自动生成

地图

描述已自动生成

文本

描述已自动生成

表格

描述已自动生成

表格

描述已自动生成

图表, 折线图

描述已自动生成

图表, 折线图

描述已自动生成

图表, 折线图, 直方图

描述已自动生成

Original Linear Model

图表, 直方图

描述已自动生成

图表, 折线图

描述已自动生成

Ridge Model

图表, 折线图

描述已自动生成

图表, 直方图

描述已自动生成

图表, 折线图

描述已自动生成

Initial RF

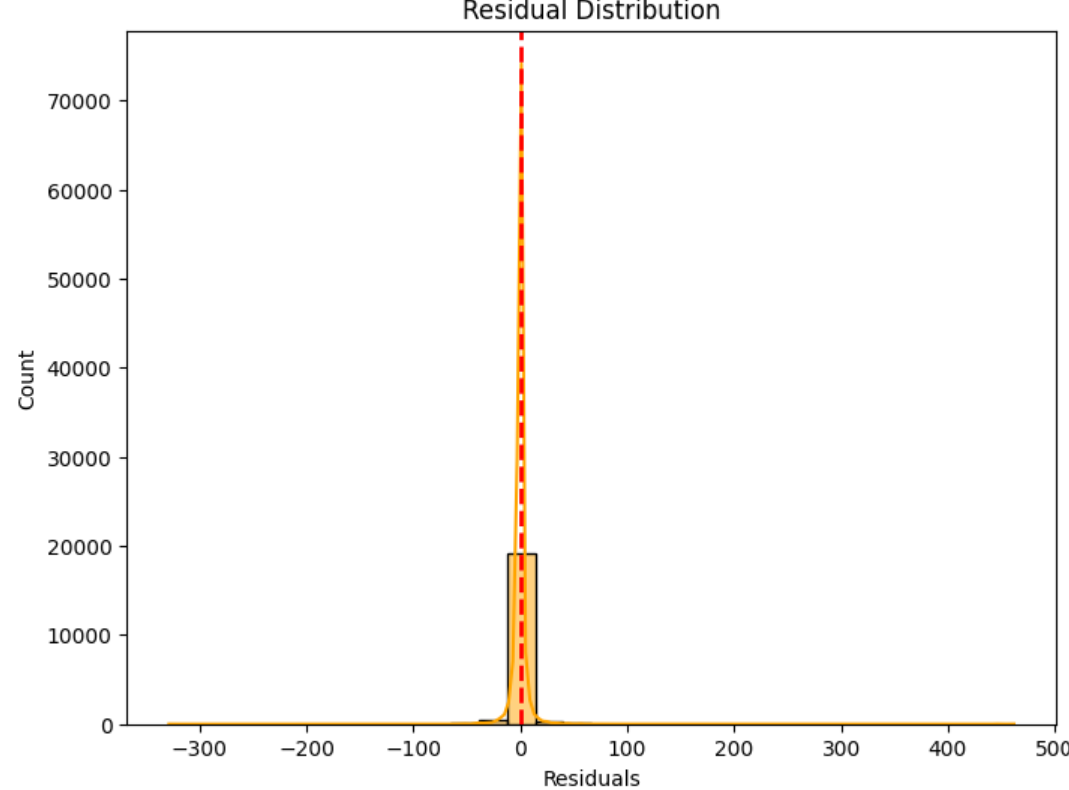
形状

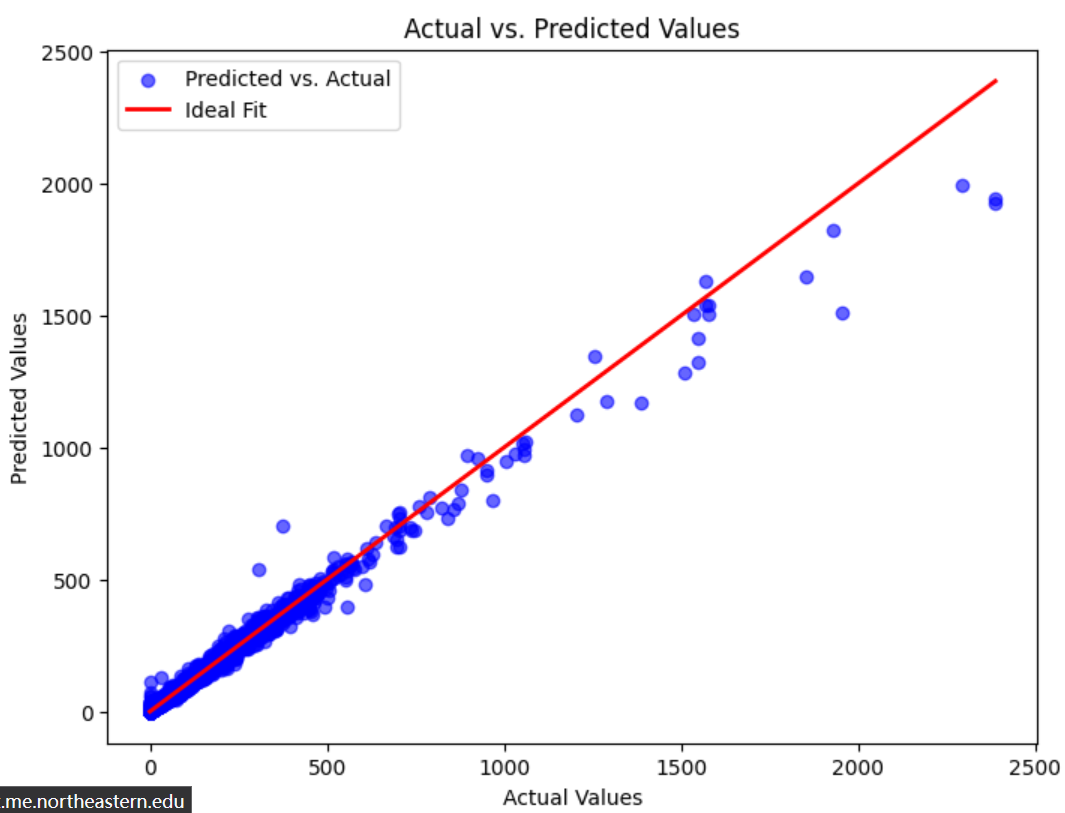
中度可信度描述已自动生成

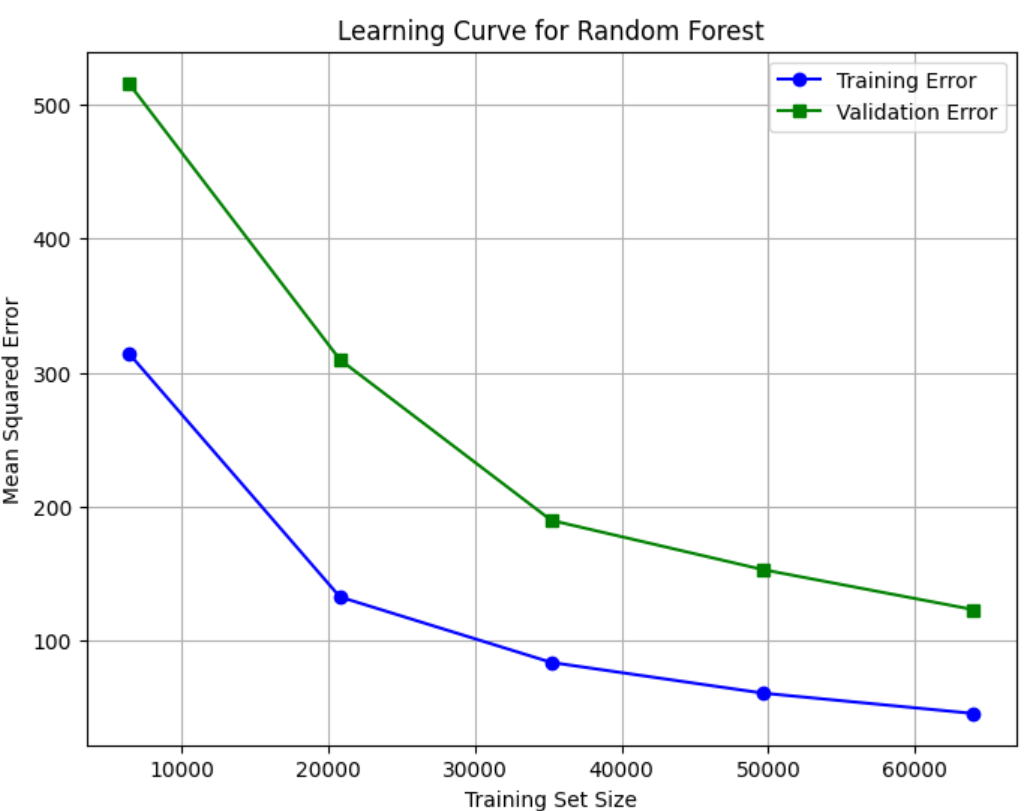
图表, 折线图

描述已自动生成

Updated RF







Initial GBM

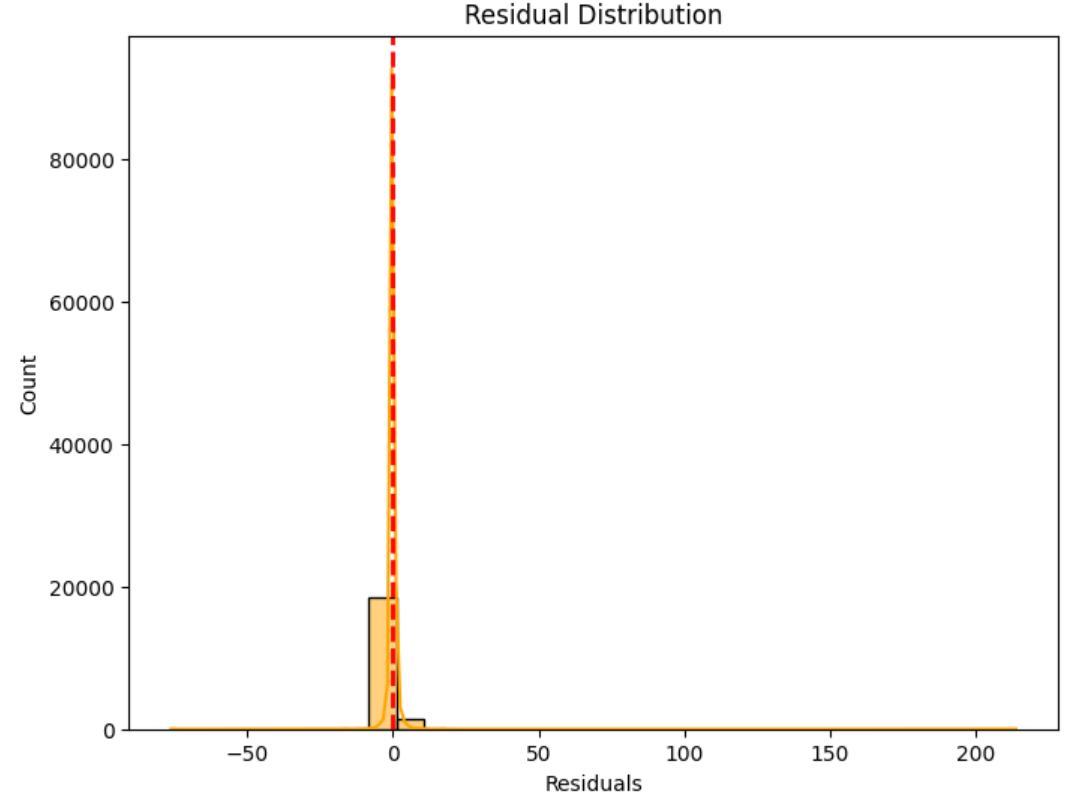
图片包含 图表

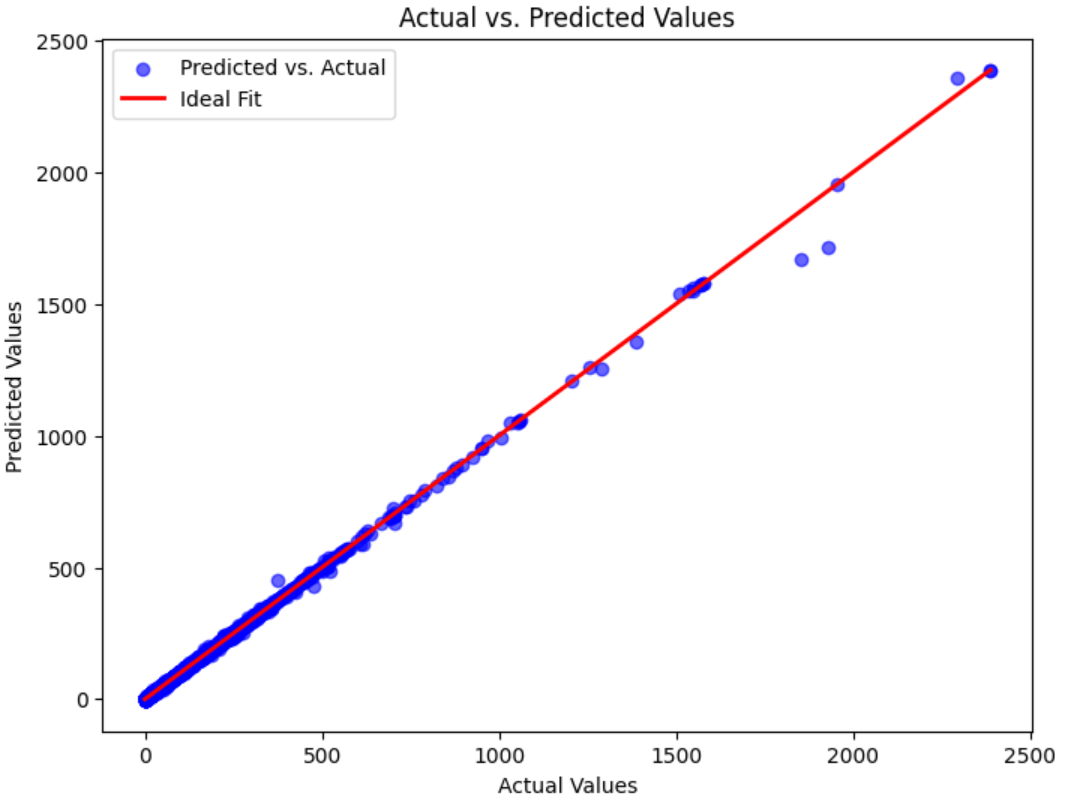
描述已自动生成

图表, 折线图

描述已自动生成

Improved GBM





图表, 折线图

描述已自动生成