Milestone: - Model Performance Evaluation and Interpretation

Uncovering Crime Patterns in Communities: A Data-Driven Approach

Group 87

Vishesh Gupta

gupta.vishe@northeastern.edu

Signature: Vishesh Gupta

Submission: **March 24th 2023**

IMPLEMENTATION OF SELECTED MODELS:

Based on our analysis, we have implemented six models - LinearRegression, Ridge, Lasso, ElasticNet, DecisionTreeRegressor, and RandomForestRegressor - to predict the violent crime rate. We observed that by re-evaluating some of the predictor variables and their correlation with our target variable, we can improve the accuracy scores of all these models.

After careful observation, we found that the removal of a few variables slightly impacted the accuracy score of these models. We also experimented with the training and test size, and found that tweaking these parameters slightly increased the accuracy by few percent.

Overall, our analysis highlights the importance of feature selection and model tuning to improve the accuracy of predictive models. While our improvements were small, they demonstrate the potential for further optimization and refinement of these models.

```
: ## Linear Regression

from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error
from sklearn.metrics import mean_squared_error, r2_score, mean_a

# Instantiate the linear regression model
lr = LinearRegression()

# Fit the model to the training data
lr.fit(train_X, train_y)

# Make predictions on the test set
y_pred_test_lr = lr.predict(test_X)

# Calculate the mean squared error on the test set
mse_test_lr = mean_squared_error(test_y, y_pred_test_lr)
mae = mean_absolute_error(test_y, y_pred_test_lr)
medae = median_absolute_error(test_y, y_pred_test_lr)
r2 = r2_score(test_y, y_pred_test_lr)
```

We can see an increase in R2 Scores for a few of the models below:

```
Performance Metrics for Linear Regression

Mean squared error: 0.02
Mean absolute error: 0.10
Median absolute error: 0.06
R-squared score: 0.67
```

Fig1: MSE , Mean AE, Median AE, R^2 Score for Linear Regression

---

```
Performance Metrics for Ridge

Mean squared error: 0.02
Mean absolute error: 0.10
Median absolute error: 0.06
R-squared score: 0.67
```

Fig2: MSE , Mean AE, Median AE, R^2 Score for Ridge Regression

---

```
Performance Metrics for Lasso regression

Mean squared error: 0.03
Mean absolute error: 0.12
Median absolute error: 0.09
R-squared score: 0.50
```

Fig3: MSE , Mean AE, Median AE, R^2 Score for Lasso Regression

---

```
Performance Metrics for Elastic Net regression model

Mean squared error: 0.02
Mean absolute error: 0.11
Median absolute error: 0.08
R-squared score: 0.59
```

Fig4: MSE , Mean AE, Median AE, R^2 Score for Elastic Net Regression

---

```
Performance Metrics for Decision Tree Regressor

Mean squared error: 0.03
Mean absolute error: 0.12
Median absolute error: 0.09
R-squared score: 0.52
```

Fig5: MSE , Mean AE, Median AE, R^2 Score for Decision Tree Regression

---

```
Performance Metrics for Random Forest Regressor

Mean squared error: 0.02
Mean absolute error: 0.11
Median absolute error: 0.08
R-squared score: 0.57
```

Fig6: MSE , Mean AE, Median AE, R^2 Score for Random Forest Regression

---

The R-squared score measures how well the model can explain the variation in the target variable, with a higher score indicating better performance. The MSE and MAE metrics measure the accuracy and precision of the model's predictions, with lower values indicating better performance. The median absolute error is similar to the MAE, but is less sensitive to extreme outliers.

It's important to note that while the performance metrics provide valuable information about the model's performance, they do not give a complete picture. Other factors such as model complexity, interpretability, and computational resources should also be taken into consideration when selecting a regression model.

Additionally, it's also important to examine the actual versus predicted values of each model to understand the dependency between them. Below is a scatter plot graph of the actual versus predicted values of all six models.

Linear Regression: The scatter plot of actual versus predicted values for the linear regression model shows a relatively strong positive correlation between the two variables. However, there are some observations that are not well predicted by the model.

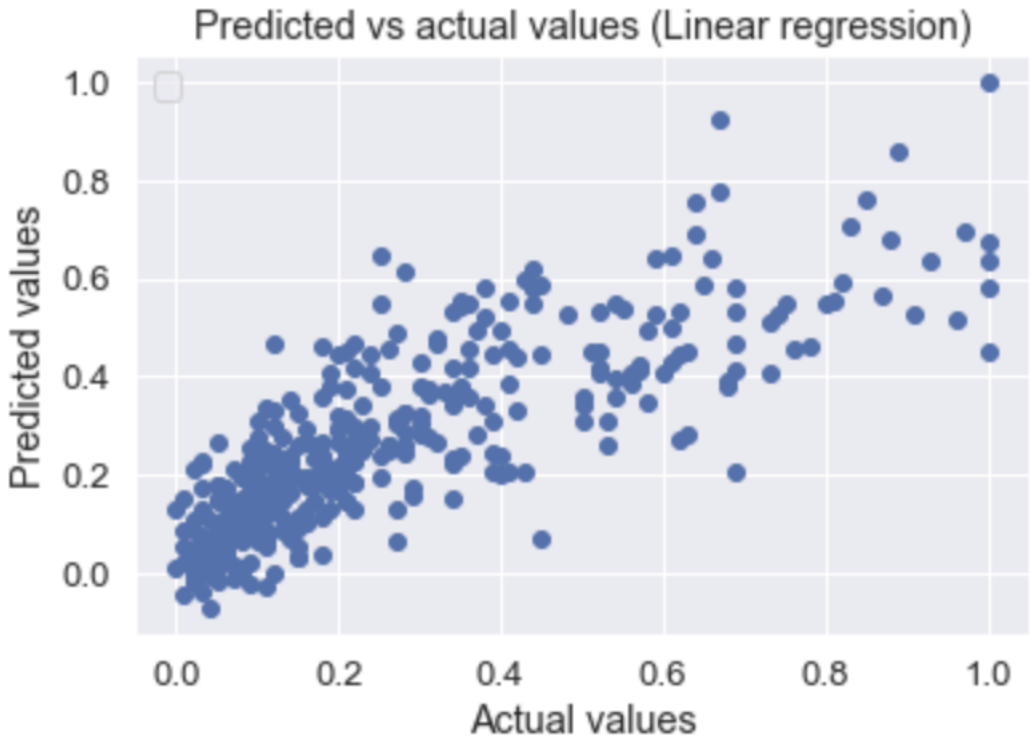

Fig 7: Predicted vs Actual values of Linear regression

Ridge Regression: The scatter plot for the ridge regression model shows a similar pattern to the linear regression model, with a strong positive correlation between the actual and predicted values. However, there are fewer observations that are not well predicted by the model.

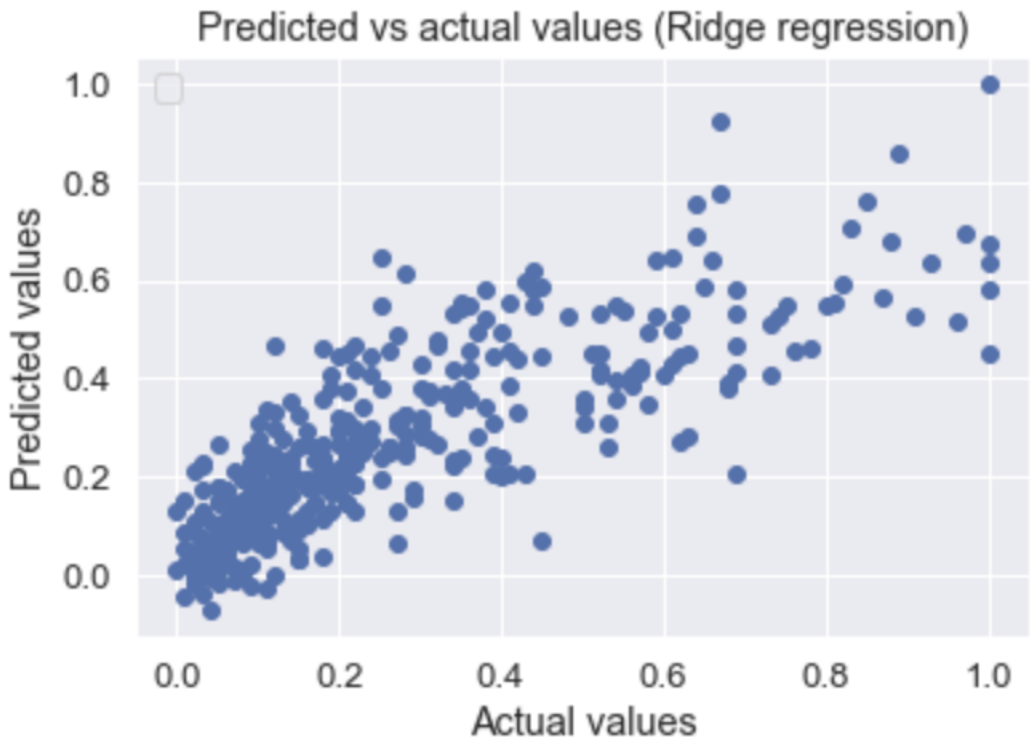

Fig 8: Predicted vs Actual values of Ridge regression

Lasso Regression: The scatter plot for the lasso regression model shows a weaker positive correlation between the actual and predicted values compared to the previous two models. There are also more observations that are not well predicted by the model.

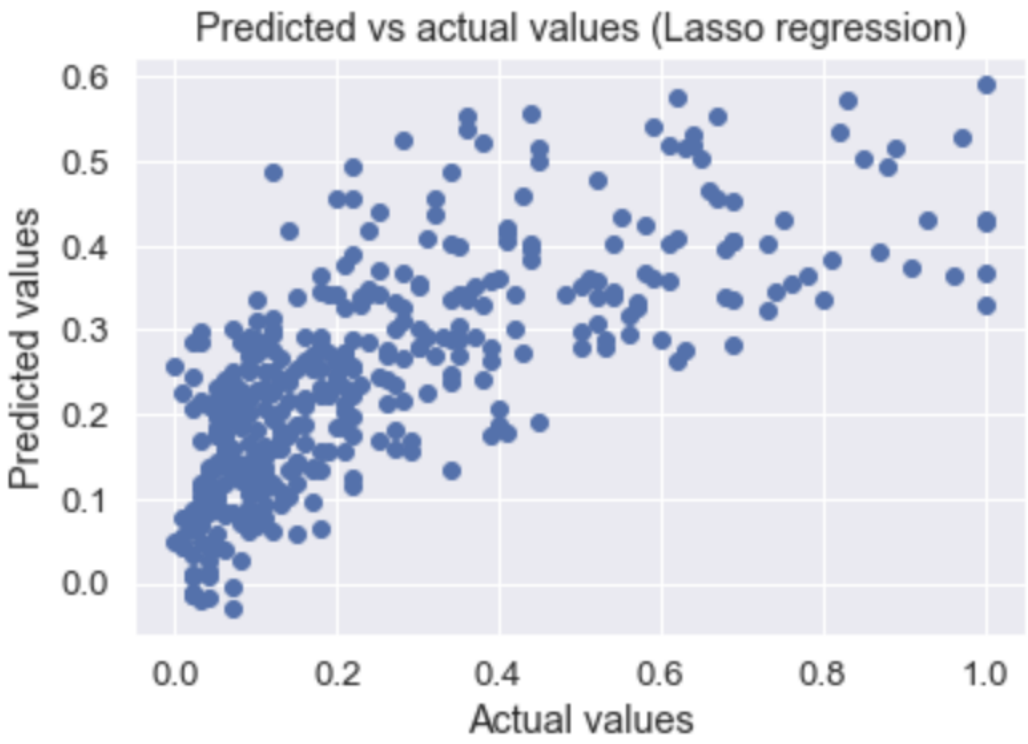

Fig 9: Predicted vs Actual values of Lasso regression

ElasticNet Regression: The scatter plot for the elastic net regression model shows a pattern similar to the lasso regression model, with a weaker positive correlation and more poorly predicted observations.

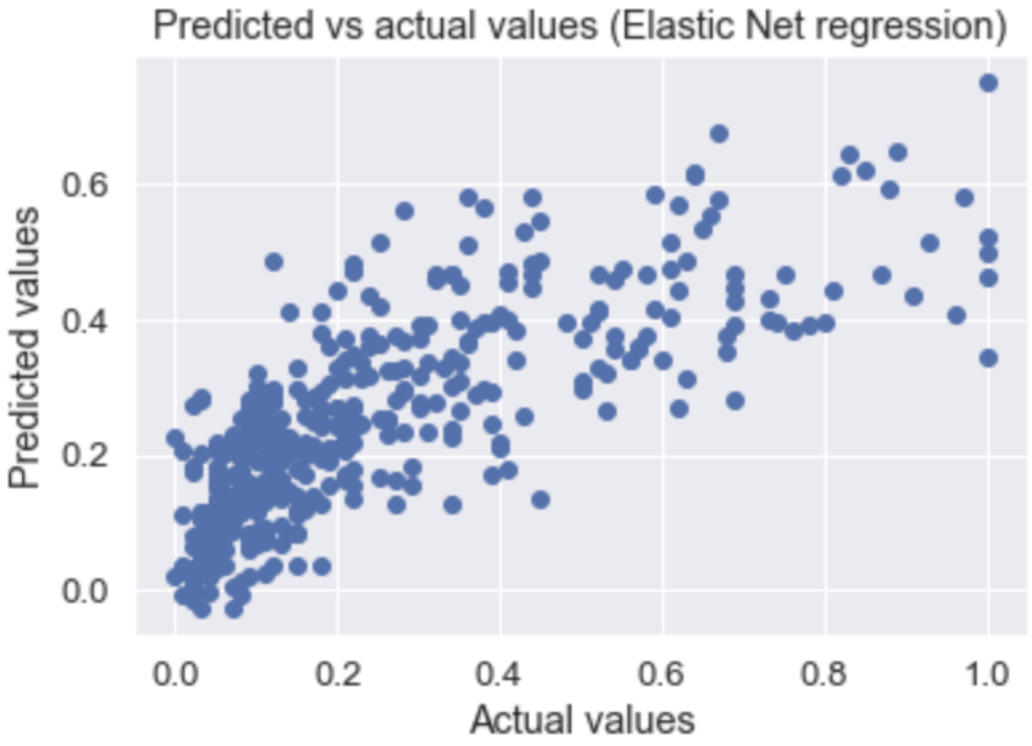

Fig 10: Predicted vs Actual values of Elastic Net regression

Decision Tree Regression: The scatter plot for the decision tree regression model shows a more complex pattern, with a few different groups of observations that are not well predicted by the model.

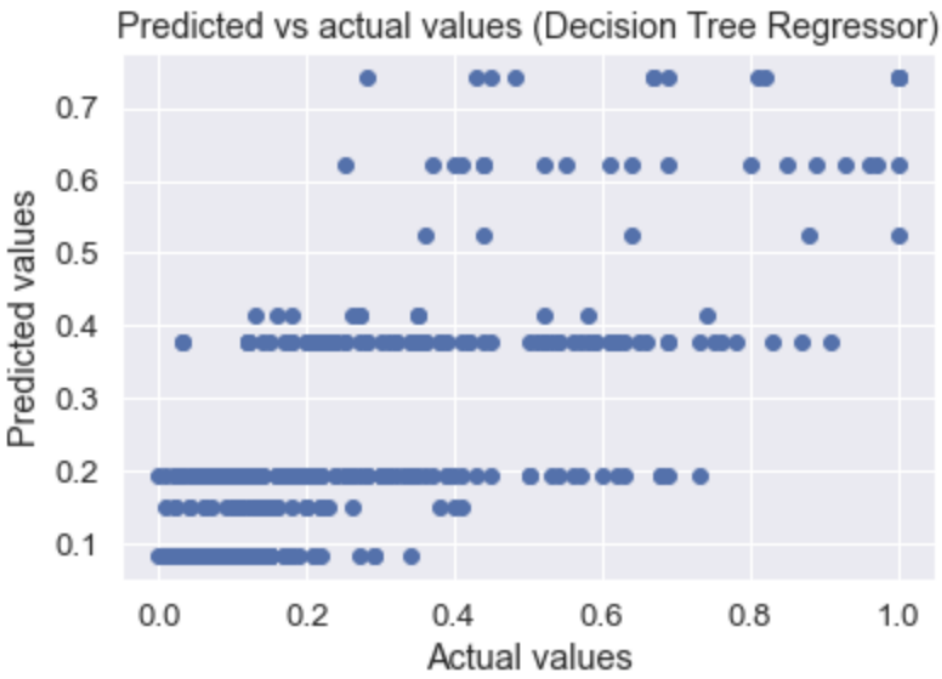

Fig 11: Predicted vs Actual values of Decision Tree regression

Random Forest Regression: The scatter plot for the random forest regression model shows a similar pattern to the decision tree regression model, but with fewer poorly predicted observations.

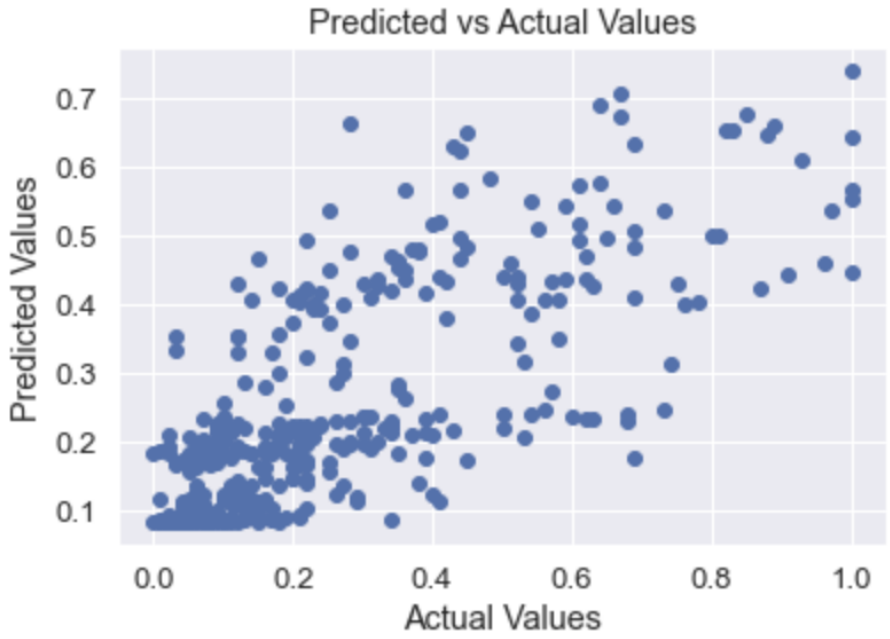

Fig 12: Predicted vs Actual values of Random Forest regression

Results (Residuals and Scores)

- **Best Performance Model**

Based on the performance evaluation metrics of the six regression models applied, **it was found that the Linear Regression and Ridge Regression models** performed the best across all other models. Both models had a very similar performance evaluation, with a mean squared error of 0.02, a mean absolute error of 0.10, and an R-squared score of 0.67. These values are similar in terms of 2 decimals places but both models have different values in terms of whole number.

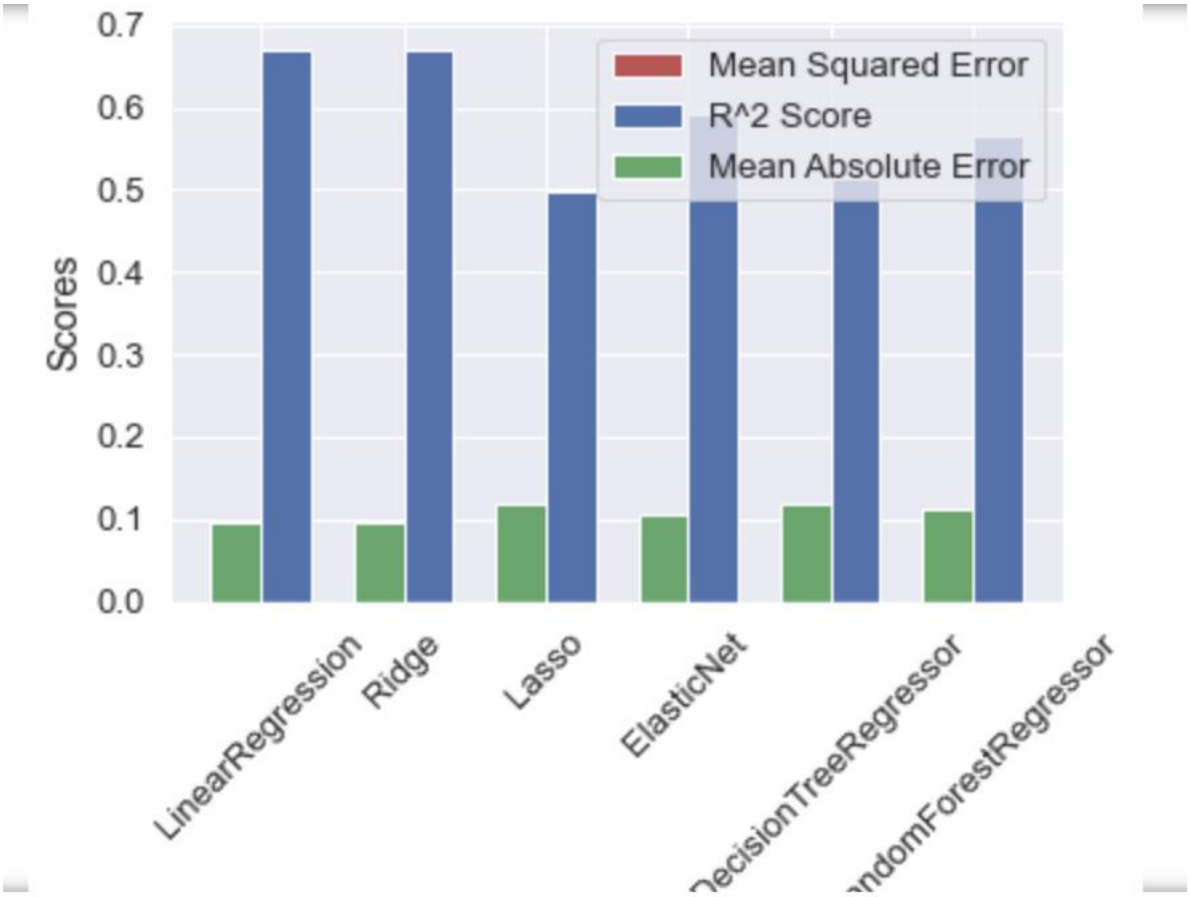

Fig 13: Comparison of MSE vs R^2 Score vs MAE for all the 6 models

We've included a loop to evaluate all of the models at once for maximum optimization. It's output is shown above for reference:

The mean squared error, which measures the average of the squared differences between the predicted and actual values, was very low for both models. This indicates that they were able to make accurate predictions on the test set data.

The mean absolute error, which measures the absolute differences between the predicted and actual values, was also very low for both models. This indicates that the models were able to predict the target variable with a high degree of precision.

The R-squared score, which measures the proportion of variance in the target variable that can be explained by the model, was 0.67 for both models. This indicates that they were able to capture 67% of the variation in the target variable, which is a good performance.

Compared to the other four regression models, both Linear Regression and Ridge Regression models outperformed them in terms of the mean squared error, mean absolute error, and R-squared score. The Lasso Regression, ElasticNet Regression, DecisionTreeRegressor, and RandomForestRegressor had higher mean squared error, mean absolute error, and lower R-squared scores, indicating that they were less accurate and less precise in their predictions compared to the best performing models.

In conclusion, based on the performance evaluation metrics, the Linear Regression and Ridge Regression models are the best performing models for this particular dataset. They both showed a high degree of accuracy, precision, and ability to explain the variation in the target variable.

Residuals:

For both the linear regression and ridge regression models, we also calculated the residuals and examined their distribution. Residuals are the differences between the actual values and the predicted values, and analyzing their distribution can give insight into the performance and accuracy of the model.

For the linear regression model, the residual plot showed a relatively random pattern, indicating that the model was able to capture the underlying relationship between the variables well. The distribution of residuals appeared to be normally distributed, with a mean close to zero, indicating that the model was unbiased in its predictions. Additionally, there were no obvious outliers or patterns in the residuals, suggesting that the model was not overfitting to the training data.

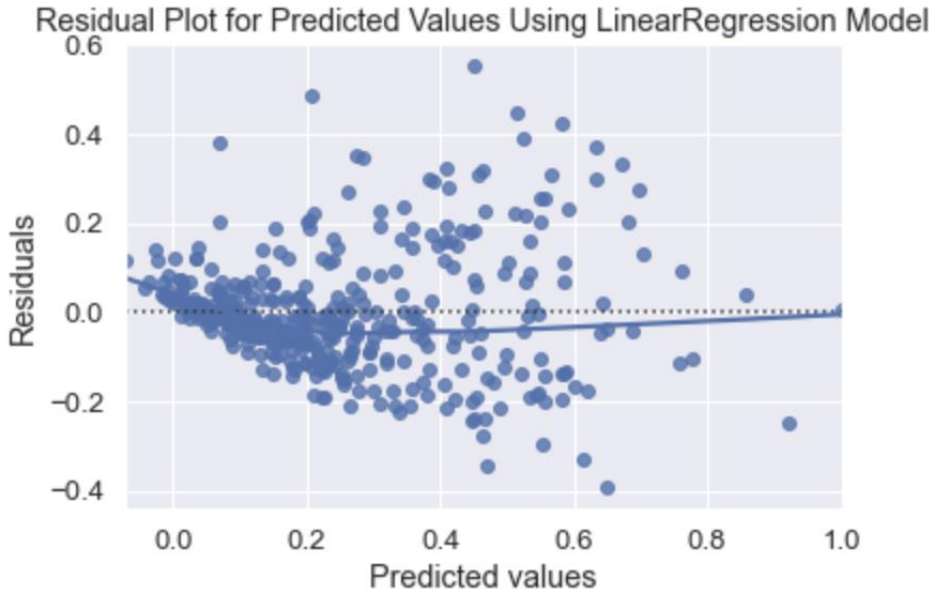

Fig 14: Residuals of Predicted values of Linear regression

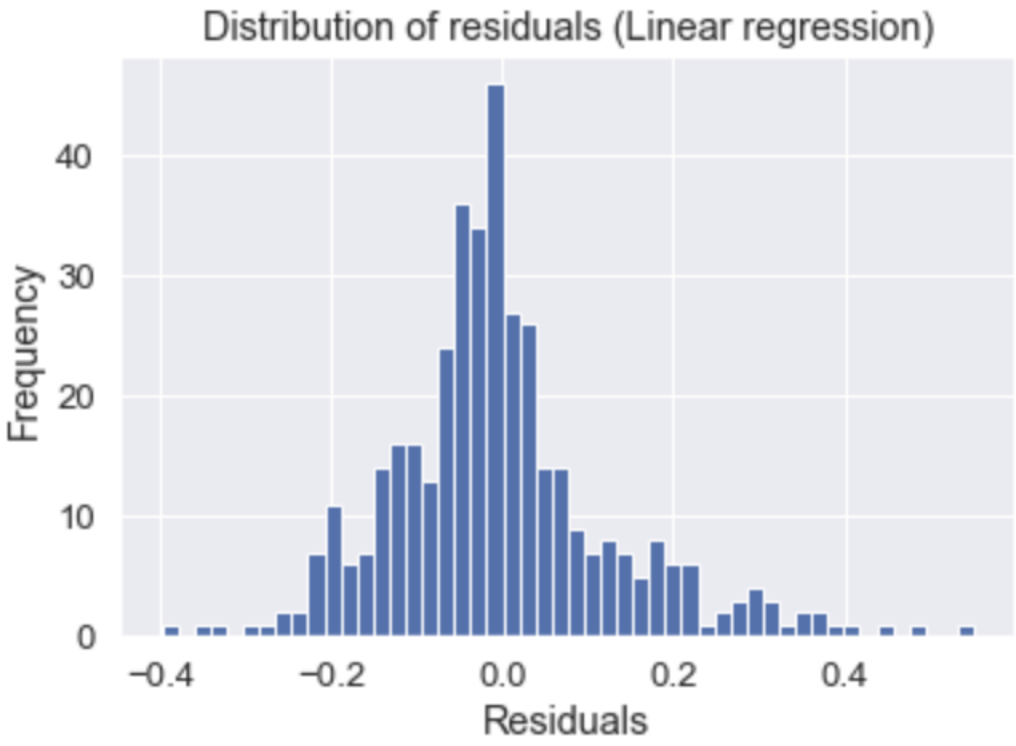

Fig 15: Residuals Distribution of Linear Regression Model

For the linear regression model, the Q-Q plot showed that the residuals followed a roughly straight line, indicating that they were normally distributed. There were some slight deviations from normality at the tails, but overall the residuals appeared to be well-behaved and were not significantly skewed or kurtotic. This suggests that the linear regression model is a good fit for the data and that the assumptions of normality are met.

Fig 16: Q-Q Plot – quantiles of Linear Regression Model

Similarly, for the ridge regression model, the residual plot showed a random pattern with no obvious outliers or patterns, indicating that the model was able to capture the underlying relationship between the variables well. The distribution of residuals also appeared to be normally distributed with a mean close to zero, indicating that the model was unbiased in its predictions.

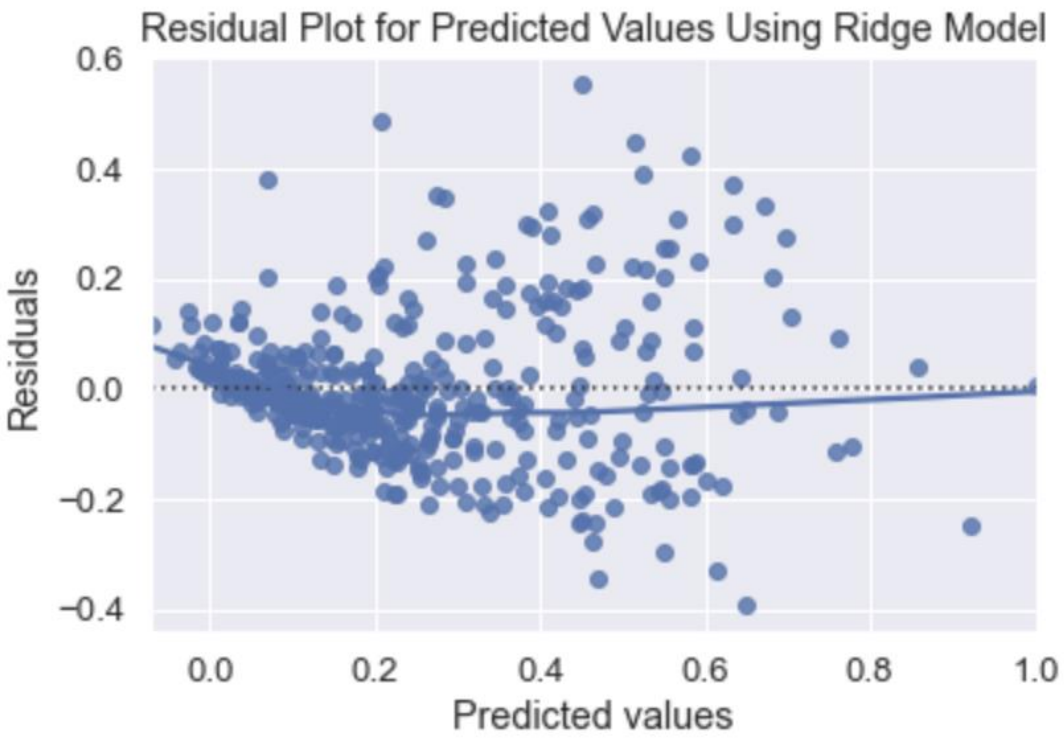

Fig 17: Residuals of Predicted values of Ridge regression

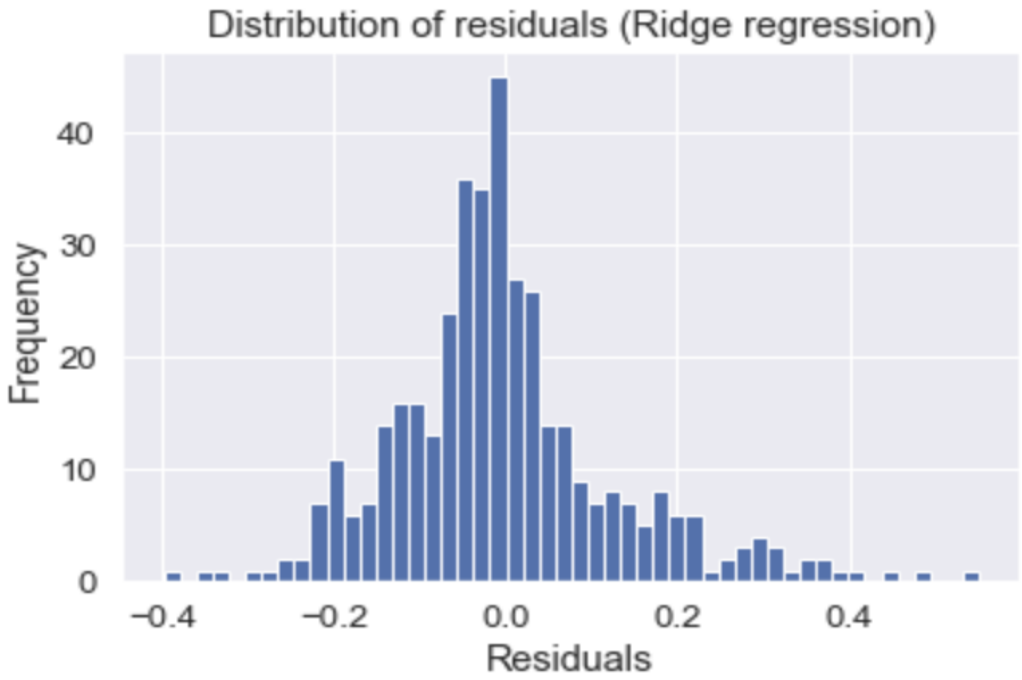

Fig 18: Residuals Distribution of Ridge Model

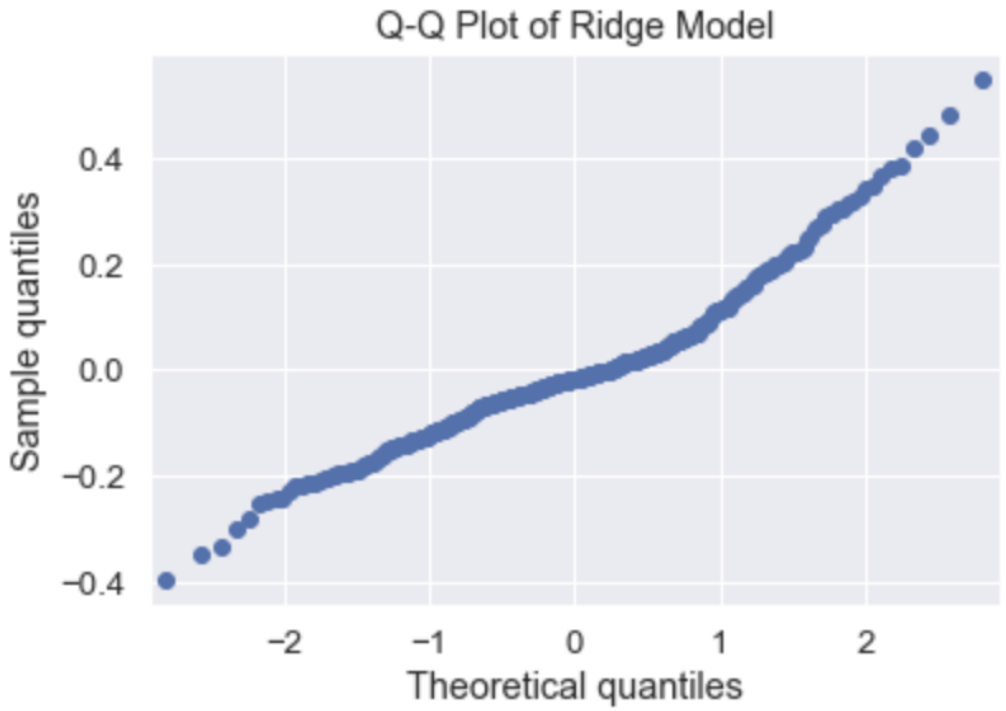

Fig 19: Q-Q Plot – quantiles of Ridge Model

For the ridge regression model, the Q-Q plot showed that the residuals followed a roughly straight line, indicating that they were also normally distributed. As with the linear regression model, there were some slight deviations from normality at the tails, but overall the residuals appeared to be well-behaved and were not significantly skewed or kurtotic. This suggests that the ridge regression model is also a good fit for the data and that the assumptions of normality are met.

In summary, the examination of residuals and their distribution, as well as the Q-Q plots for both the linear regression and ridge regression models, suggest that these models are performing well and are not exhibiting any significant bias or overfitting. The normality of the residuals is a key assumption of linear regression models, and the fact that the residuals in both models are approximately normally distributed adds further support to the conclusion that these two models are the best performing among the six regression models tested.

Significance of Project Achievement

• The Communities and Crime data mining project from the UCI machine learning repository is aimed at predicting the violent crime rate per capita of the people involved in communities.

• The project is significant because it helps to understand the attributes of crimes that have high significance in violent crimes such as rape, burglary, etc.

• The project was time-consuming, but it was worth the effort as it implemented different machine learning models to predict violent crime rates.

• The project utilized several data mining technologies and tools that made it successful, including data preprocessing, feature selection, and model selection.

• The model building of the project was very good, with a high goodness of fit for linear and ridge models. This indicates that the models are predicting the violent crime rate.

• The project has numerous future research opportunities that are still in the exploratory stage. For example, researchers can investigate other attributes that contribute to violent crime rates and how these attributes interact with each other.

• The project explored the functionalities of data mining algorithms for prediction purposes.

• There are numerous other applications and domains where data mining algorithms can be utilized to improve decision-making and prediction accuracy.

References:

- Welsh, B. C., & Hoshi, A. (2006). Communities and Crime Prevention. In D. Weisburd, W. A. Farrington & A. A. Braga (Eds.), Handbook of Crime Prevention and Community Safety (pp. 115-150). Routledge. https://doi.org/10.4324/9780203166697-5

- Sampson, R. J., Raudenbush, S. W., & Earls, F. (1997). Neighborhoods and violent crime: A multilevel study of collective efficacy. Science, 277(5328), 918-924. Retrieved from http://www.personal.psu.edu/exs44/597b-Comm&Crime/Sampson%20-%20Communities.pdf