Can violent crimes in a community be predicted beforehand and prevented? A deep dive into the UCI Communities and Crime Dataset.

Debangshu Bhattacharya

December 15, 2020

Abstract

Every time we see the news of violent crimes in our television sets, we feel more and more concerned and disheartened. The question ultimately arises whether this could have been somehow predicted beforehand and hence prevented. With this very question in mind, this project aims to study the Communities and Crime dataset [3] available publicly in the UCI Machine Learning Repository to try to find some insightful details and some possible explanatory variables that can predict such atrocities. The ultimate aim of the model created is to predict the number of violent crimes(murder, rape, robbery, and assault) per 100K population of a given community. The paper revolves around finding a set of important features (actual or engineered) and building a regression model which satisfies all the model assumptions. Our Ridge regression model with MAE value of 0.0482 on the chosen set of important features not only outperform the MAE metric of 0.096 recorded in [4], but also, we can draw a confidence interval around our predictions to reduce uncertainty.

Keywords: Crime prediction, Regression, Machine Learning

1 Introduction

There has been various studies in crime prediction literature to find potential explanatory variables and crime patterns[2]. Most of the work done in the literature revolves around predicting zones or communities into the respective class of crime rate. Some classify the zones/areas into the classes 'high', 'medium' or 'low' crime rate zones [1] while some treat it as a binary classification problem of classifying into either 'high' or 'low' crime rate zone[5] [6]. In this paper, we treat the problem as a **regression problem** instead and we predict the target variable of "ViolentCrimePerPop" which is the number of violent crimes per 100K population scaled to the range 0-1. Redmond et.al [4] used various Case-Editing Approaches to remove potential noisy instances and recorded a best Mean Absolute Error of 0.096 for a 10-fold cross validation of the data. In this paper, we focus on finding the most important features which can explain the variability of the target variable values instead. We further refine the feature space by removing the features which contribute the most towards multi-collinearity. Finally, we use a Ridge Regression model(as there is still some multi-collinearity among chosen features) for our dataset and observe that our model does significantly better with a Mean absolute error of 0.0482 for a 10-fold cross validation of the data.

The various sections in the paper are given as follows:

- Section 2.1 gives a description of the dataset used.
- Section 2.2 presents the Exploratory data analysis on the training dataset and how the features are selected for the model.
- Section 2.3 presents the model methodology.
- Section 2.4 specifies how the hyper-parameter of the model was tuned to have an optimal model. Further a comparison of the train and test set metrics of the given model are provided.
- Section 2.5 checks if the assumptions of Linear Regression are valid for the model chosen.

- Section 2.6 compares the metrics for some other Machine Learning models for which we lose out on explainability.

2 Database and Methodology

2.1 Problem Statement

The dataset that used in the project is the “Communities and Crime Dataset” available publicly in UCI Machine learning repository. The link for the dataset is as follows: `https://archive.ics.uci.edu/ml/datasets/Communities+and+Crime+Unnormalized`. The data contains **2215 instances and 147 attributes**. The dataset comprises real values of communities in the United States and has been compiled from socio-economic data from the 1990 US Census, law enforcement data from the 1990 US LEMAS survey, and crime data from the 1995 FBI UCR. There are 18 possible goal variables out of which I will be trying to predict the number of violent crimes per 100K population. The other 17 goal variables will be dropped. For the target variable, there were 221 instances which had missing values. So these instances were dropped. The target variable is converted to a scale of 0-1. We explicitly keep apart a separate test set which consists of 200 examples. All our exploratory data analysis and model selection is performed on the remaining 1794 instances. After fixing the model, we evaluate the model on the test set. We also compare the performance of the model with other baseline metrics in available research literature. For this, the 10-fold cross validated mean score of the whole data is observed and compared.

2.2 Exploratory Data Analysis and Feature Selection

From figure 1 we can see the distribution of our original target variable and also the Normal Q-Q Plot of the same. We can clearly see that the target variable is not gaussian. We also performed the Kolmogorov Smirnov test for the same and the null hypothesis of normality is rejected. We will transform our target variable such that it is normally distributed. For this, we use the Box-Cox transformation as:

$$\hat{y} = \frac{(y+\delta)^\lambda - 1}{\lambda}$$

where $\delta = 10^{-16}$ is used such that $(y+\delta)$ is positive and as Box-Cox transformation is applied on positive values. The optimum value of $\lambda = 0.17198268$ is chosen that makes $\hat{y}$ normal. The transformed data and the corressponding Normal Q-Q plot are shown in figure 2. A Kolmogorov Smirnov test for normality was also performed on $\hat{y}$ and the null hypothesis of normality could not be rejected.

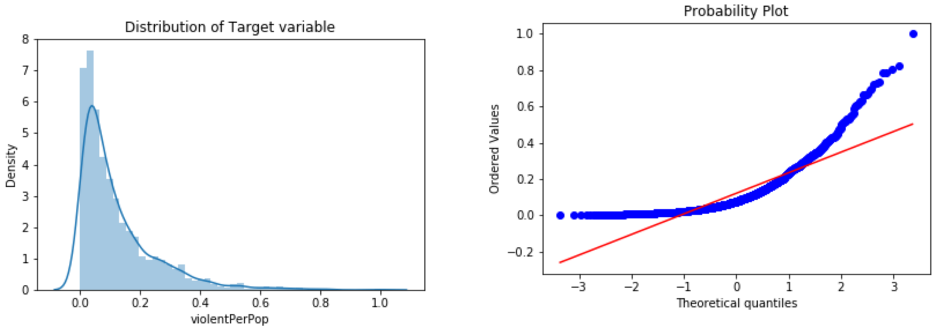


Figure 1: *Part (a): Left Subplot shows the distribution of the original target variable with ’violentPerPop’ variable values(scaled to 0-1) in the X-axis and the Kernel Density estimate in the Y-axis. Part (b): Right Subplot shows the Normal Q-Q plot of the original target variable with the theoretical Quantiles in X-axis and Observed Quantiles in Y-axis. We can clearly see from these two plots that the original target variable is not normally distributed and hence we need some transformation.*

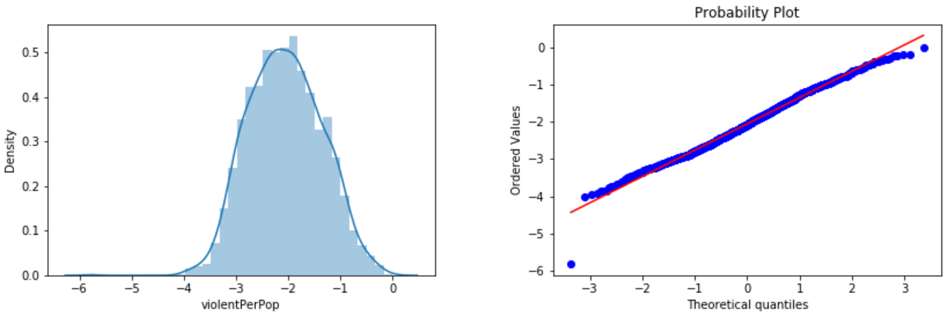


Figure 2: *Part (a): Left Subplot shows the distribution of the Box-Cox tranformed target variable with the tranfored target variable values in the X-axis and the Kernel Density estimate in the Y-axis. Part (b): Right Subplot shows the Normal Q-Q plot of the transformed target variable with the theoretical Quantiles in X-axis and Observed Quantiles in Y-axis. Graphically, the tranformed target variable seems to be normally distributed with a possible outlier in the left tail.*

Then, for each feature we check the percentage of missing values. We see that there are 22 variables which have more than 84% missing values and for 2 variables there are around 59% missing values. We drop these variables from consideration. There is also another variable which has around 5% missing values. We impute the missing values of this variable as the median value of that variable. Further, we drop the variables (‘State’ and ‘communityname’) as these are informative variables about the community and should not considered as features in model building. The important features are obtained from the remaining 103 variables.

For initial feature selection we perform the following steps:

1. We look at the histogram of the features. This gives us an idea about possible transformations on the features which may be needed due to presence of outliers or a wide range of values.
2. We also look at the scatter plot of the feature with the target variable and check if a linear association is visible. If there is a quadratic or polynomial association we also generate artificial features to capture those trends.
3. We perform a mutual information score for each feature with the target variable. If the mutual information score is above 20% we consider that feature for further steps else we drop it. In figure 3, we see the top 10 important features among the first 30 variables.

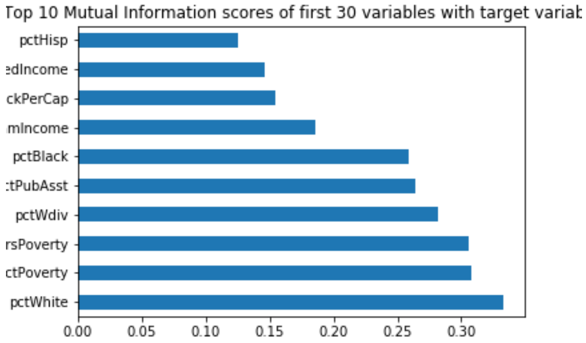


Figure 3: *From the first 30 variables the mutual information scores are calculated for each variable with respect to the transformed target variable. The top 10 variables and their corresponding mutual information scores are shown in the figure above.*

For feature transformation for some of the features, a log transformation is used. The reasoning and intuition behind using a log transformation can be seen from figure 4. In figure 4:

- We see in the upper left image the distribution of the original variable 'persPoverty'. We can clearly see that the wide range of values and possible outliers are having a significant impact.
- In the upper right image, the original variable 'persPoverty' is plotted against the transformed target variable. We do not see much association between the variables. However, from the mutual information score (figure 3), we know that 'persPoverty' is an important feature. Since, the wide range of values is the main issue, we take a log transform on the feature as $\hat{x} = c * log(x + d)$ where we chose the values c = 1 and d = 15.
- The bottom left image is the density plot of the log transformed feature 'persPoverty'.
- The bottom right image is the scatterplot of the transformed feature and transformed target variable. Here, we can clearly see the linear association between the two.

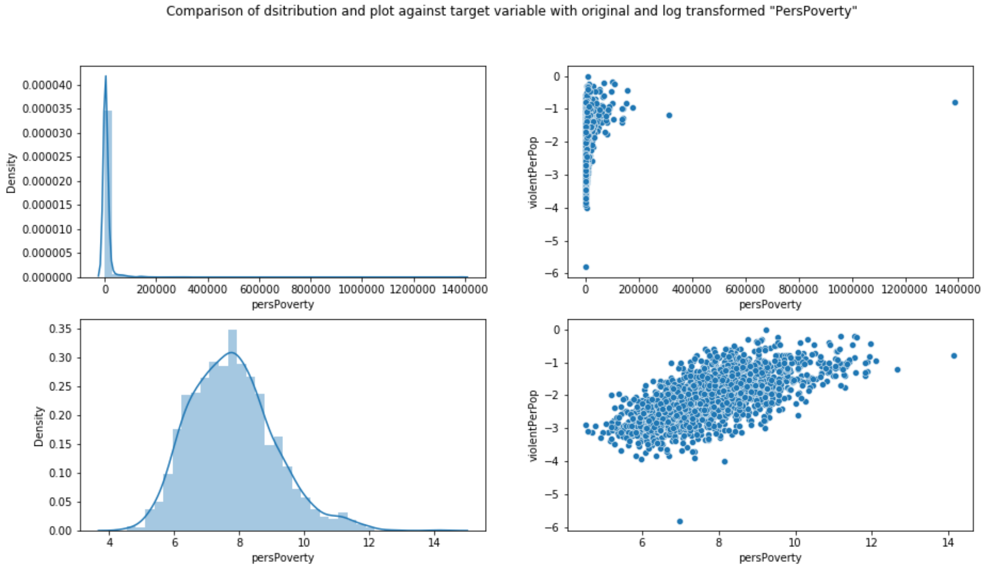


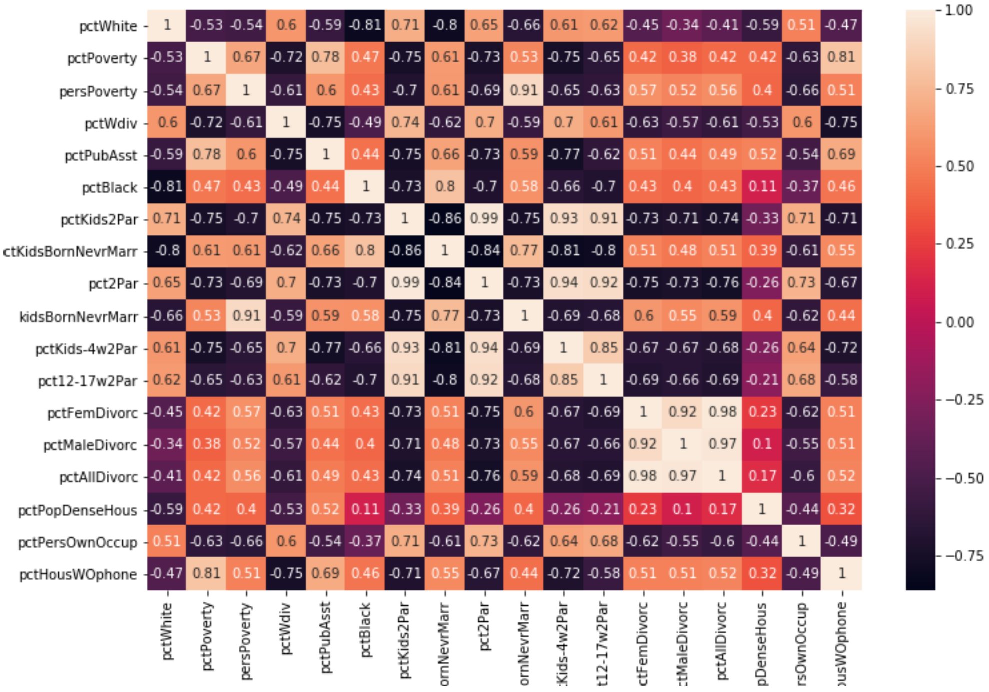
Figure 4: *a) Upper left subplot shows the original kernel density plot of the feature 'persPoverty' b) Upper right subplot shows the scatterplot between the original feature values of 'persPoverty' with the transformed target variable. We can clearly see from this plot that the scale of the original feature variable and outliers are causing a lot of problem and we cannot see any visible pattern in spite of the fact that this is an important feature. c) Lower left subplot shows the kernel density plot of the same feature with a log transform. d) Lower right subplot shows the scatterplot between the transformed feature and transformed target variable.*

From the initial feature selection, we get a total of 18 features. Next we check for the multi-collinearity among these features. We see from the correlation heatmap plot (figure 5 and the Variation Inflation Factor values that there are some features which are highly correlated with each other. For a pair of features which have high correlation between them, we remove the feature with the higher VIF. Proceeding this way, we remove 8 features till we are left with 10 features. The corresponding VIFs for these 10 features are shown in table 1. We can see that there are still some features with high VIF. We observed that removing these features hamper the efficiency of our model. So we are going to keep these features. But as the features suffer from multi-collinearity we are going to use Ridge Regression instead to deal with this problem.

The selected 10 features and their plot against the transformed target variable is shown in figure 6. From the plot, we can see that a qudratic association with the transformed target variable is present for the features 'pctPoverty','pctBlack','pctPopDenseHous', 'pctKidsBornNevrMarr'. So, 4 artificial features are created which are just the square of these respective features.

Feature	VIF
pctPoverty	13.41
pctWdiv	32.11
pctPubAsst	12.68
pctBlack	05.41
pctKidsBornNevrMarr	12.92
kidsBornNevrMarr	48.82
pctMaleDivorc	20.35
pctPopDenseHous	02.68
pctPersOwnOccup	35.58
pctHousWOphone	08.73

Table 1: Variation Inflation Factor of 10 features selected



	pctWhite	pctPoverty	persPoverty	pctWdiv	pctPubAsst	pctBlack	pctKids2Par	ornNevrMarr	pct2Par	ornNevrMarr	:Kids-4w2Par	t12-17w2Par	ctFemDivorc	:tMaleDivorc	pctAllDivorc	pDenseHous	rsOwnOccup	ousWOphone
pctWhite	1	-0.53	-0.54	0.6	-0.59	-0.81	0.71	-0.8	0.65	-0.66	0.61	0.62	-0.45	-0.34	-0.41	-0.59	0.51	-0.47
pctPoverty	-0.53	1	0.67	-0.72	0.78	0.47	-0.75	0.61	-0.73	0.53	-0.75	-0.65	0.42	0.38	0.42	0.42	-0.63	0.81
persPoverty	-0.54	0.67	1	-0.61	0.6	0.43	-0.7	0.61	-0.69	0.91	-0.65	-0.63	0.57	0.52	0.56	0.4	-0.66	0.51
pctWdiv	0.6	-0.72	-0.61	1	-0.75	-0.49	0.74	-0.62	0.7	-0.59	0.7	0.61	-0.63	-0.57	-0.61	-0.53	0.6	-0.75
pctPubAsst	-0.59	0.78	0.6	-0.75	1	0.44	-0.75	0.66	-0.73	0.59	-0.77	-0.62	0.51	0.44	0.49	0.52	-0.54	0.69
pctBlack	-0.81	0.47	0.43	-0.49	0.44	1	-0.73	0.8	-0.7	0.58	-0.66	-0.7	0.43	0.4	0.43	0.11	-0.37	0.46
pctKids2Par	0.71	-0.75	-0.7	0.74	-0.75	-0.73	1	-0.86	0.99	-0.75	0.93	0.91	-0.73	-0.71	-0.74	-0.33	0.71	-0.71
ctKidsBornNevrMarr	-0.8	0.61	0.61	-0.62	0.66	0.8	-0.86	1	-0.84	0.77	-0.81	-0.8	0.51	0.48	0.51	0.39	-0.61	0.55
pct2Par	0.65	-0.73	-0.69	0.7	-0.73	-0.7	0.99	-0.84	1	-0.73	0.94	0.92	-0.75	-0.73	-0.76	-0.26	0.73	-0.67
kidsBornNevrMarr	-0.66	0.53	0.91	-0.59	0.59	0.58	-0.75	0.77	-0.73	1	-0.69	-0.68	0.6	0.55	0.59	0.4	-0.62	0.44
pctKids-4w2Par	0.61	-0.75	-0.65	0.7	-0.77	-0.66	0.93	-0.81	0.94	-0.69	1	0.85	-0.67	-0.67	-0.68	-0.26	0.64	-0.72
pct12-17w2Par	0.62	-0.65	-0.63	0.61	-0.62	-0.7	0.91	-0.8	0.92	-0.68	0.85	1	-0.69	-0.66	-0.69	-0.21	0.68	-0.58
pctFemDivorc	-0.45	0.42	0.57	-0.63	0.51	0.43	-0.73	0.51	-0.75	0.6	-0.67	-0.69	1	0.92	0.98	0.23	-0.62	0.51
pctMaleDivorc	-0.34	0.38	0.52	-0.57	0.44	0.4	-0.71	0.48	-0.73	0.55	-0.67	-0.66	0.92	1	0.97	0.1	-0.55	0.51
pctAllDivorc	-0.41	0.42	0.56	-0.61	0.49	0.43	-0.74	0.51	-0.76	0.59	-0.68	-0.69	0.98	0.97	1	0.17	-0.6	0.52
pctPopDenseHous	-0.59	0.42	0.4	-0.53	0.52	0.11	-0.33	0.39	-0.26	0.4	-0.26	-0.21	0.23	0.1	0.17	1	-0.44	0.32
pctPersOwnOccup	0.51	-0.63	-0.66	0.6	-0.54	-0.37	0.71	-0.61	0.73	-0.62	0.64	0.68	-0.62	-0.55	-0.6	-0.44	1	-0.49
pctHousWOphone	-0.47	0.81	0.51	-0.75	0.69	0.46	-0.71	0.55	-0.67	0.44	-0.72	-0.58	0.51	0.51	0.52	0.32	-0.49	1

Figure 5: *The correlation heatmap of the 18 most important features initially selected is shown in the figure. We can easily see that there is a huge multi-collinearity problem. There are some groups of features which are highly correlated with each other and from these groups we can drop all but one feature. For example the three features 'pctFemDivorc', 'pctMaleDiv' and 'pctAllDivorc' are highly correlated and we can just select one out of these 3 features instead for model building.*

2.3 Model Methodology

In this project, I am exclusively working on this dataset as a **regression problem** modelling the number of violent crime per population(scaled to 0-1) as the target variable. A major part of the problem was to select the most important features among the possible 103 numerical features (There were a total of 129 features but 24 of these features had a significant amount of missing values while 2 features are non predictive. So, these 26 features had to be dropped from consideration). We have successfully selected the 10 most important features and engineered 4 extra features, reducing the feature space to a 14 dimensional vector. The next part of the problem is to select the best model. We not only want a regression model which can minimize the generalization error but also can model the uncertainty. For

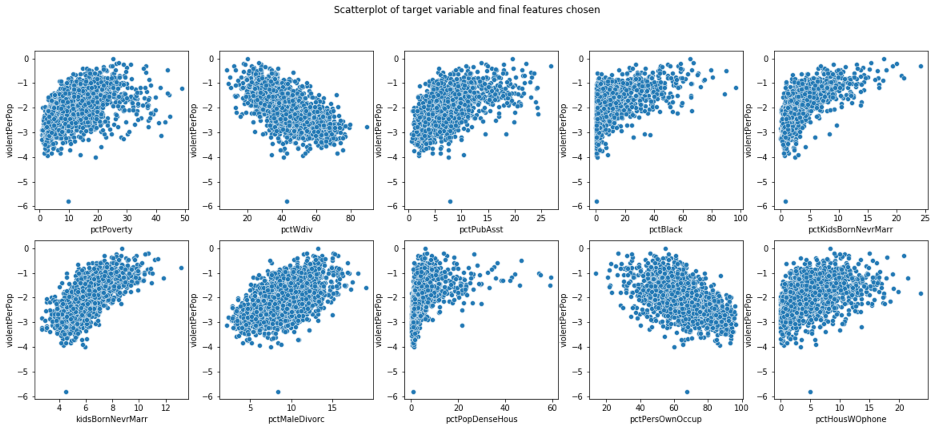


Figure 6: Scatter plot of the features with transformed target variable

this, we would prefer simpler models rather than complicated models which lose out on explainability. For example, in Table 3, the 10 fold cross validation metrics for different models are shown which shows that a Neural Network model is performing the best. However, as we do not have the power of explainability in a NN model, we consider the next best model which is a Ridge Regression model.

2.4 Outlier Removal, Model Selection and Evaluation

2.4.1 Outlier Removal

For each of the feature variables and the goal variable, the Inter Quartile Range (IQR) value was computed. The lower limit(l) and upper limit(u) of acceptance was set as $l = Q1 - 4.5 * IQR$ and $u = Q3 + 4.5 * IQR$, where Q1 and Q3 are the first and third quartile values of that variable. If a value(x) is less than l or greater than u, we say that x is an outlier. 73 such instances were found and dropped from the training dataset. The resultant training dataset is of size 1721 instances.

2.4.2 Model Selection

For model selection, the optimal regularization parameter lambda for ridge regression was chosen by the parameter value which gave minimum cross validated mean squared error over lambda $\in (0, 10)$. The optimum value of lambda chosen is 7.898. Now, the selected model is fitted over the outlier removed train data which consists of 1721 instances with our selected 14 features, scaled by Standard Normalization, and transformed target variable $\hat{y}$.

2.4.3 Evaluation

For evaluation, the model predicts the value $\hat{z}$ for which the original value is y. The value $\hat{z}$ is transformed to z as follows:

$$z = e^{\frac{\ln(\lambda\hat{y}+1)}{\lambda}} - \delta$$

Now the metrics Root Mean Squared Error, Adjusted R2-Score and Mean Absolute Error are computed on the true values(y) and predicted values(z). This evaluation is performed both on the training set and the test set of 200 instances. The values observed are shown in table 2.

	Train	Test
RMSE	0.074	0.065
MAE	0.046	0.043
Adj.R2	0.581	0.624

Table 2: Model Evaluation on train and test set

2.5 Model Diagnostics

Since the model chosen is a Ridge Regression model and the problem is a Linear Regression problem, we check if the assumptions of Linear Regression are satisfied by the model.

1. **Linearity:** For checking the assumption of linearity we plot the fitted values against the model residuals. The plot obtained is shown in 7. The correlation coefficient between fitted and residual values was also obtained and is very low (-0.005). From the graph in figure 7, we cannot conclude any clear patterns and hence the assumption of linearity is justified.

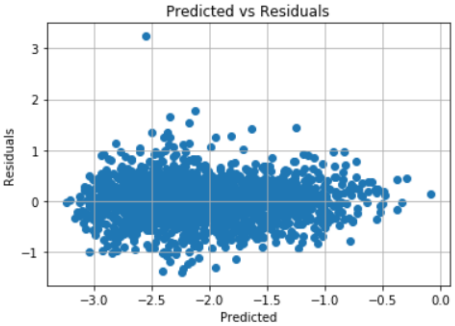


Figure 7: Fitted vs Residuals

2. **Randomness:** For checking the randomness of the residuals, the 'randtest' package in Python was used and the assumption of randomness is justified.

3. **Homoskedasticity:** For checking the homoskedasticity of residuals the Bartlett test was performed and with a 5% significance level we cannot reject the null hypothesis of homoskedasticity (p value = 0.6)

4. **Normality of residuals:** For checking the assumption of normality of residuals we plot the Normal Q-Q plot (figure 8) of the residuals and also perform the Kolmogorov Smirnov test. We also cannot reject the null hypothesis of normality for a 5% significance level (p value = 0.15).

Thus all the assumptions of Linear Regression model is satisfied.

2.6 Model Comparison

Now, as the model satisfies all of its assumptions and it is doing reasonably well, we fix the model. We now compare this model with some other regression models by performing a 10 fold cross validation on the whole data and comparing the metrics Mean Squared Error(MSE), Mean Absolute Error(MAE) and R2 score. The comparison of the metrics is shown in table 3.

We can see that the Neural Network model (trained on all 102 numeric features) is outperforming all the other regression models. We have not done any explicit feature selection but used 2 hidden layers of 500 and 100 nodes respectively and the output layer contains one node. Also, as the Support Vector Regressor, Random Forest Regressor and Neural Network Regressor models are not hyper-parameter tuned, the metrics can probably be even further improved upon tuning.

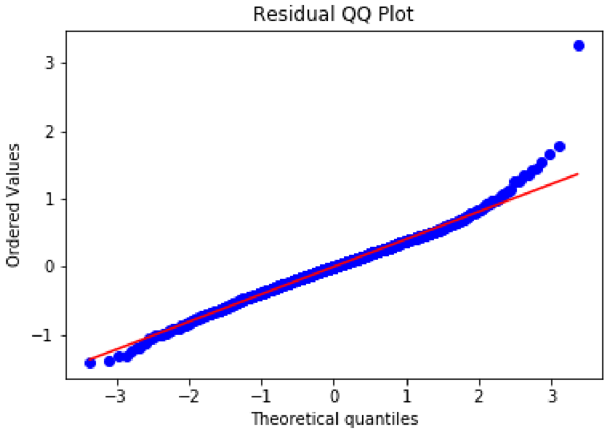


Figure 8: Residuals Q-Q Plot

However, given that the Ridge Regression model is:

1. a simple model for which we do not lose out on explainability.
2. able to draw confidence intervals around the predictions to remove uncertainity.
3. having metrics not too worse off than the best model metrics.

we can fairly be certain that our model is indeed doing pretty well.

Model	**MSE**	**MAE**	**R2**
Ridge Regression	0.0060	0.0482	0.6130
Support Vector Regression	0.0061	0.0483	0.6118
Random Forest Regression	0.0063	0.0485	0.5966
Neural Network Regression	**0.0058**	**0.0458**	**0.6326**

Table 3: Model Comparison

3 Conclusion

We started the project with the serious question of whether violent crimes can be predicted or not. From this case study of the UCI communities and crime dataset, we can certainly say that even a simple Ridge regression model can learn the crime pattern and predict with a certain confidence the amount of violent crimes that may happen in a given community. This can greatly be used as a flagging mechanism and government bodies can further employ this to strengthen security measures for those regions. Now, there is no reason that the crime pattern in a certain region (USA in our example) will be representative of the true crime pattern of the whole world. However, local government bodies can adopt the methodology to their own private datasets. The model can be further enhanced if time series data of the crime rates are provided for a specific region. To conclude, crime prediction is certainly achievable and can be used extensively by local governments as a flagging mechanism to strengthen security and measures.

4 Future Works

For future works, one can extend the question of whether crime can be predicted to another important question of whether there is any inherent bias in the crime prediction model. Ideally we don't want any inherent bias in our model and there is a wide area of research on how to identify and remove such bias from our model. This is another aspect I would like to work on in the future provided the opportunity.

References

[1] Syahid Anuar, Ali Selamat, and Roselina Sallehuddin. Hybrid artificial neural network with artificial bee colony algorithm for crime classification. In *Computational Intelligence in Information Systems*, pages 31–40. Springer, 2015.

[2] Anna L Buczak and Christopher M Gifford. Fuzzy association rule mining for community crime pattern discovery. In *ACM SIGKDD Workshop on Intelligence and Security Informatics*, pages 1–10, 2010.

[3] M Redmond. Communities and crime unnormalized data set. *UCI Machine Learning Repository. In website: http://www. ics. uci. edu/mlearn/MLRepository. html*, 2011.

[4] Michael A Redmond and Timothy Highley. Empirical analysis of case-editing approaches for numeric prediction. In *Innovations in Computing Sciences and Software Engineering*, pages 79–84. Springer, 2010.

[5] Bharti Suri, Manoj Kumar, et al. Performance evaluation of data mining techniques. In *Information and Communication Technology for Sustainable Development*, pages 375–383. Springer, 2018.

[6] Prajakta Yerpude. Predictive modelling of crime data set using data mining. *International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol*, 7, 2020.